IDENTITY OF LONG-TAIL ENTITIES IN TEXT

FILIP ILIEVSKI

The research reported in this thesis has been funded by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza fund, which was granted to prof.dr. Piek Vossen in 2013.

Dutch title: *Identiteit van laag-frequente entiteiten in tekst*

Cover: photo by `https://pixabay.com/users/comfreak-51581/` adapted by Frosina Ilievska and Filip Ilievski

Printed by: Grafoprom Bitola || `http://grafoprom.com`

Identity of Long-tail Entities in Text by *Filip Ilievski* is to be published by *IOS Press* in the book series *Studies on the Semantic Web* in 2019.

VRIJE UNIVERSITEIT

Identity of Long-Tail Entities in Text

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geesteswetenschappen
op vrijdag 6 september 2019 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door
Filip Ilievski
geboren te Bitola, Macedonië

promotoren:	prof.dr. P.T.J.M. Vossen
	prof.dr. F.A.H. van Harmelen
copromotoren:	dr. K.S. Schlobach
	dr. M.G.J. van Erp

committee:	prof.dr. L. de Vries
	dr. T. Kuhn
	dr. D. Maynard
	prof.dr. M. de Rijke
	prof.dr. W. Daelemans
	prof.dr. P. Cimiano

ACKNOWLEDGMENTS

This has been one of the most difficult sections to write. How to sufficiently express my gratitude to all people that were crucial for my PhD life over the past four years? It seems complicated to do justice here. Still, the acknowledgments are probably going to be the most visited section of this thesis, so I'd better perform well. . . :-)

The PhD years have been a great shaping experience. The intuition that I belong in academia, such a vibrant environment with fundamental discussions and critical minds, was apparently not wrong. It was a real delight to spend days with inspiring people, and a worthwhile challenge to practice structuring my quite-limitless freedom that came with the PhD. I did not mind either traveling to remote places to discuss ideas with colleagues from worldwide. All in all, this has been an exciting life chapter.

Above all, I must acknowledge the role of my mentors. I have been extremely lucky to have you guide me through my PhD years, and I believe we made a great team! I was very much inspired and challenged by each one of you. Your directness and care fed my motivation throughout the process. Besides being great academic minds, your views and interests set an example in a more general, human sense - which I find equally important. Thank you for all of this!

Piek, thanks for your incredible devotion and guidance at every stage of my PhD process. Seeing the energy you share with your students and staff, I am still wondering whether you have found a trick to expand your day beyond the customary 24 hours. I could not have wished for a more careful, reliable, and understanding promotor. I appreciate greatly that you stand firmly for what you believe and find sufficient strength to voice your opinion on important societal matters. More selfishly, thanks for largely shaping my PhD topic, for generating brilliant ideas, and for allowing me to visit a number of conferences and spend months at CMU.

Frank, your passion for science and our discussions during my Master studies gave me a first taste of what research is like, which I did not resist. Thanks for being the prime inspiration for me to direct myself towards the academia. Your rigorous and fair take on science is something I greatly appreciate. I am also amazed to date that you can always summarize my thoughts better than I can. Thanks for the interesting discussions, and for never losing track of the broad picture of my research, which, I believe, helped me tremendously in writing the thesis without much hassle.

Stefan, because of the *systematicity* in writing I learned from you, I will never again write a paper without listing my hypotheses early in a table :-p Your continuum between a mentor, a collaborator, and a friend has been very comfortable, and at times refreshingly challenging. Starting from my exchange year at the VU in 2012/13, throughout my Master studies, Master thesis, and the PhD period,

I have always felt your strong faith in my abilities and your thoughts about my best interests. I would like to thank you for your huge patience and empathy in my worst moments, and for your appreciation and support in my best ones. The valuable pieces of advice I received during our meetings will long fuel my goals in the future.

Marieke, I would like to thank you for introducing me to the academic culture, for patiently helping me understand the research practices, and your assistance in learning LaTeX. Your role in facilitating various papers in the early stages of my PhD was a major boost of my self-confidence and a waiver of many potential worries later. Besides your support and guidance, thank you for being probably the most cheerful and easygoing researcher I know, always lighting up the mood around you, and constructively looking for ways forward. I still intend to watch Star Wars, by the way, it's on my list ;)

Ed, I would like to express my gratitude for your hospitality at Carnegie Mellon in 2017, which was an extremely rich and impactful experience for me. The ideas we were passionate about then are still at the heart of my research interest today. Thank you for your guidance and eye-opening suggestions, especially in Santa Fe last year: I have never understood my strengths and weak points clearer than on that day. I am still fascinated with the very detailed emails we exchange at the expenses of other (I am sure, more important) tasks of yours. You are a great motivation and a source of inspiration for me to date.

Marten, it has been a pleasure to grow together for so long during our PhDs, sharing many challenges, several moments of triumph, and two inspiring trips to America with you, Chantal, and Pia. Emiel and Marten, it was not less entertaining to inject all those song lyrics randomly in conversations. A little less conversation, little more action, please ;-) I hereby apologize to the CLTL members for any inconvenience caused, though I bet you secretly enjoyed it. . . Chantal and Pia, best of luck with completing your PhDs. Hennie, Isa, Selene, all others: thank you for the nice time spent in and around the office together. Each of you has been an amazing colleague and a very unique character.

Wouter and Laurens, I cannot imagine a more fun and fruitful start of my PhD than our collaboration and coffee breaks in T3.12. Laurens, I promise you an actual competitive tennis match in a few years. Wouter, I learned a lot from you about both science and arts; Sun Ra is still among the most unusual jazz musicians I have encountered. I hope your political ideas keep evolving and influencing the (academic) world in the next period. Albert, thank you for your friendliness and curiosity, and for being an excellent mix of science and music. Ilaria, it has been awesome to hang out both in academic as well as in social setting; looking forward to our first NLP collaboration. Erman, Anca, Annette, Peter, Victor, Jan, Veruska, and all others: it has been a delight to share the KR-R/SW group.

I could not imagine my life in Amsterdam without my social circles. Eduard, thank you for the fun chats and the cheeky insights, that might be occasionally useful. Eduard and Adrien, the Sunday evening jam sessions, though tiring, have motivated me on many Monday mornings. Laurie, you have been a great

addition to my life: I appreciate our long talks about various topics and our exploration of the city culture with you and Adrien. Thanks for your kindness and for encouraging my goals. Unmesh, thanks a lot for always being a great friend and for supporting my sporting targets. Elle, I enjoyed sharing lunches and gym sessions with you, always a pleasure to spend some time together. Alberto, Nedim, Herbi, thank you guys for your company and the parties we shared; I hate that Amsterdam has been a transit city for each of you. Andrej, Angela, Robbert, Shex, Kaja, Gabor, Laurens - it has been a great joy spending time with each one of you.

I would like to express my appreciation of my dear friends in Macedonia. Dragan, Jovan, Jasna, Marija, Matej, Deks, Maja, Kocho, Cvete, Tatar, Darko, Andrej - you have made my every trip over there worthwhile. I am fascinated with the strength of our connection in spite of the distance, and I am looking forward to many more years of socializing and creating memories.

Here is a shout-out to all my dear people spread around the globe: Dimi, Leandro, Amote, Evangelia, Andjela, Marina, Natalia, Berke - I can not wait to see you again and hear what is new.

I am hereby sending hugs to my family. Mama, tato, Frosina, my deepest bows for your unconditional support and understanding, for all your patient and sincere advice, and everything you taught me during my younger years!

Finally, I would like to thank all members of my thesis committee, for their valuable reviews, timely communication, and dedication.

Filip
Amsterdam, 2019

CONTENTS

ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
EL	Entity Linking
EnC	Entity Coreference
EvC	Event Coreference
IAA	Inter-annotator agreement
IR	Information Retrieval
KB	Knowledge Base
KG	Knowledge Graph
LDF	Linked Data Fragments
LE	Lexical Expression
LOD	Linked Open Data
NER	Named Entity Recognition
NLP	Natural Language Processing
QA	Question Answering
SRL	Semantic Role Labeling
WSD	Word Sense Disambiguation

1 INTRODUCTION

1.1 BACKGROUND: IDENTITY IN THE DIGITAL ERA

Identity is an abstract representation of an entity, consisting of the individual characteristics by which it is recognized or known.[1,2] For instance, the identity of the VU University is defined by its location (the street de Boelelaan in Amsterdam), its founding year (1880), its funding model, etc. My own identity is a combination of non-alterable inherent properties, such as birthplace and race, but also certain choices I made in life so far, such as my education. This notion of identity relates to two related, but different, notions: personal identity and sameness/equivalence.

Personal identity has been heavily discussed in the domain of philosophy, questioning the persistence of an entity (typically a person) through time. This has resulted in various ideas as basis for contrasting theories. According to the body theory (Thomson, 1997; Williams, 1957), one's identity is preserved as long as he/she exists in the same body, whereas Locke's memory theory (Locke, 1689) proposes that identity should be dependent on the extent to which one can consciously recall. Hume's Bundle theory (Hume, 1738) argues that there is no permanent self, as our bodies undergo continuous qualitative change.

Another related, but different, notion is that of equivalence or 'sameness' between real world entities, which carries the problem of no clear distinction whether two real-world entities are referentially identical. There are several dimensions to be considered here, including temporality (Is Old Amsterdam identical to New Amsterdam?), pragmatism (Is Lord Lipton identical to the wealthiest tea importer?), and granularization (Are passengers and people who traveled with EasyJet an identical set?) (Recasens et al., 2011).

Societal relevance While the philosophical aspects touching upon the notion of identity and its related concepts are important, recently identity has gained practical relevance, shifting from a philosophical to a societal issue. The digital era and the omnipresence of computer systems feature a huge amount of data on identities (profiles) of people, organizations, and other entities, in a digital format. Within a digital identity, one typically receives a unique identifier and is described through a set of attribute values. Notably, digital identities do not necessarily need to correspond to identities in the physical world: a single world object could be represented with many digital identities, and even non-existing objects can easily exist digitally.

1 http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=identity&h=0000&j=1#c

2 https://en.oxforddictionaries.com/definition/identity

This fluid nature of digital identity makes it vulnerable to malicious actions, like stealing one's personal identity (identity theft), and using that identity to commit fraud (identity fraud). Mistaken identity is responsible for a number of scandals, such as the well-known case of the former British MP, Lord Alistair McAlpine, who was falsely accused of child abuse in November 2012.[3] While identity theft and fraud is not new, its magnitude has been growing rapidly.[4] For instance, approximately 15 million Americans have their identities used fraudulently each year, leading to a $50 billion worth of losses.

Establishing and singularizing identity Provided that the architecture of the world wide web in its current form does not connect the various sources of information on a data level, users are required to create redundant profiles on different websites, which means that the current Internet features a proliferation of these profiles representing our digital identities. Notably, the problem of digital identity is far wider than the resolving of sameness between structured user profiles in the social media or in online merchant shops, since the majority of the web content today is estimated to be in unstructured format.[5] This includes various textual documents, such as news documents, encyclopedias (such as Wikipedia), personal websites, books, etc., but also different modalities, namely, videos, images, sounds. All of these contain precious descriptions of millions, or even billions, of identities in their spatio-temporal-topical context.

Given that all these pieces of information are complementary, it is crucial to be able to combine them into a single representation. Humans are very successful in extracting information from various sources and abstracting over the form to preserve the semantics of the content. While machines have the potential to automate this process and scale it far beyond human capacity, their current capabilities to extract, combine, and reason over such information are far from the desired level of accuracy. For this reason, many companies, including Facebook, Amazon, and Google, are in a race to singularize digital identities automatically across social networks, in news documents, in books, video and image collections, etc.[6]

Establishing identity from text Looking closer at the text modality, we face the same obstacles mentioned above. Depending on its relevance, an entity can appear very seldom or very often in written language, being it news documents, encyclopedias, books, tweets, or blogs. Establishing one's identity then requires efficient and effective ways to interpret language and map it to an existing representation. For this purpose, it is customary to adhere to Kripke's "direct reference" proposal and the so-called "causal theory of reference". According to this position, echoed later by Tim Berners-Lee as one of the key principles of the Semantic Web, every thing gets "baptized" with a unique name which refers, via

3 https://www.adelaidenow.com.au/news/world/bbc-director-general-george-entwistle-quits-after-mp-falsely-accused-of-child-sex-abuse/news-story/e75d213a629c990072ceb6966aaf0d4a?from=public_rss

4 https://en.wikipedia.org/wiki/Identity_theft#Identity_cloning_and_concealment

5 http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

6 https://adexchanger.com/ecommerce-2/why-amazon-facebook-are-in-hot-pursuit-of-digital-identity/

some causal chain, directly to its referent. This allows the referents of a name to be transmitted unambiguously through time, space, and other possible worlds. Baptized names can be associated with a description, however, descriptions are not synonymous with the name and do not need to be strictly satisfied when identity is established (Halpin, 2012).

Challenges There are several key reasons why identity extraction and singularization from text is challenging for machines. From the philosophical perspective, the difficulty relates to the non-trivial question of equivalence between two descriptions of an entity, mentioned above. Integration of information is an under-addressed challenge too, as most AI and NLP research at the moment concerns isolated, local tasks. Another set of challenges stems from the inherent pragmatics of language, and its dependence on context and background knowledge - these are discussed next.

Time-bound complexity of human language Semantic processing defines a relation between natural language and a representation of a world it refers to. A challenging property of natural language is the time-bound complex interaction between lexical expressions (LEs) and world meanings. We use *meaning* as an umbrella term for both concepts and (event and entity) instances, and *lexical expression* as a common term for both lemmas and surface forms. We can define this interaction as a set of relations, both sense relations and referential relations, that exists within a language community in a certain period of time, e.g., one or a few generations. The people belonging to these generations share one language system that changes relatively slowly but during their lives there are many rapidly changing situations in the world that make certain meanings and expressions dominant and others not. For instance, while the expression 'Tesla' nowadays dominantly refers to the car company or a specific car model, several decades ago this meaning did not exist in the world and the default interpretation of 'Tesla' was the famous inventor, *Nikola Tesla*.

Likewise, we expect that a generation uses a certain set of lexical expressions out of the available set in relation to a set of meanings that balances the trade-off between learning many expressions and resolving extreme ambiguity of a small set of expressions. For this purpose, when referring to the famous inventor, *Nikola Tesla*, nowadays, one would typically use his full name, because the dominant meaning of the form 'Tesla' is the car company (unless, of course, the reference of this form to the inventor is made clear through the surrounding context). The aforementioned trade-off between efficiency and effectiveness is dictated by pragmatic principles of human language, the discussion on which comes next.

The efficiency of human language Besides time, our use of language is also dependent on other contextual aspects, including our social relationships, a topic, a location, and a community. Textual documents are surrounded by such rich context that is typically leveraged by humans but largely ignored by machines. Ambiguity of language resolves using this context: people optimize their communication to convey maximum information with minimum effort given the specific situation. Regardless of the genre (newswire, tweets, fiction, etc.), the

Gricean maxim of quantity (Grice, 1975) dictates that an author makes assumptions about the familiarity of the reader with the events and entities that are described in a document at the time of publishing. The author uses this to formulate a message in the most efficient and scarce, yet understandable way. The reader is expected to adequately disambiguate forms and fill in the gaps with presumed knowledge from the current world.

For example, when reading a news item, human readers are aware on which date it was published, which events occurred around that date, which entities are in the news, and what are the recent news articles. Machines, on the contrary, are deprived of such context and expectations. They usually have to deal with individual, isolated news articles, and need to establish identity solely on the basis of a single document in relation to dominant entities in the available resources. To overcome this, we need to build algorithms that can fill contextual knowledge gaps similar to humans with the right assumptions on the familiarity of the entities within a given context. These considerations are particularly relevant for **long-tail entities** that are only known within very specific contextual conditions. The long-tail entities are described in the next section.

1.2 CHALLENGE: ENTITY LINKING IN THE LONG TAIL

It is common to establish identity of entities in text by **Entity Linking**. The task of Entity Linking (EL) anchors recognized entity mentions in text to their semantic representation, thus establishing identity and facilitating the exploitation of background knowledge, easy integration, and comparison and reuse of systems.[7] Current EL datasets and systems typically perform linking to existing entity representations in Wikipedia[8] and its structured correspondents, DBpedia[9] and Wikidata[10].[11] For instance, in the sentence 'John Smith currently has five self-released albums out', "John Smith" refers to `http://dbpedia.org/resource/John_Smith_(musician)`. By establishing this interpretation relation, we immediately have access to pre-existing structured knowledge about the interpreted entity, such as its gender, place of birth, or education. At the same time, certain background knowledge about the existing entity representation is essential in the first place, in order for one (a machine or a human) to be able to decide between this musician and the hundreds of other politicians, artists, and sportsmen that share the same label. Hence, background knowledge has a double role in this process: it helps disambiguation and it enriches the knowledge about the

This discussion is based on the position paper: Filip Ilievski, Piek Vossen, and Marieke van Erp (2017). "Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking." In: *International Conference on Language, Data and Knowledge*. Springer, Cham, pp. 143–149

7 In this thesis, the terms 'mention', 'name', 'surface form', 'form', and 'label' are used interchangeably to refer to a proper noun phrase that appears in a document and refers to an entity instance.

8 `http://wikipedia.org`

9 `http://dbpedia.org`

10 `http://wikidata.org`

11 Throughout this dissertation, the term 'dataset' is synonymous with an NLP evaluation dataset, i.e., a collection of text documents with ground-truth annotations.

entity behind the form. These two roles complement each other and gradually increase the quality and quantity of the knowledge.

Ambiguity and variance As in this example, often a number of entity instances share the same surface form.[12] A surface form that can refer to M different entities is *ambiguous* with an ambiguity degree of M. Conversely, an entity instance can be referred to by one out of N forms, e.g., John Smith would sometimes be called 'John', 'Mr. Smith', 'singer', 'musician', 'the white man', etc. The amount of different surface forms that refer to an entity constitutes its *variance*. The ambiguity of surface forms and the variance of entity instances follow the very elegant interplay of human language with pragmatics, described through the Gricean maxims.

While humans mostly do not even perceive the ambiguity and variance of language, this M-to-N mapping between surface forms and instances, in combination with the fluid contextual relevance of those instances, makes the task of entity linking challenging for machines. Nevertheless, state-of-the-art EL systems (Cheng and Roth, 2013; Daiber et al., 2013; Moro et al., 2014; Nguyen et al., 2016a; Usbeck et al., 2014; Zwicklbauer et al., 2016) report high accuracies, which seemingly rejects the impression that these systems struggle with capturing the pragmatics of language with its ambiguity and variance.[13] What exactly is happening here?

Long-tail entities In this thesis, I hypothesize that these accuracies are largely due to the dominance of a limited number of popular entities. The accuracy of most probabilistic algorithms is mainly based on test cases for which there is sufficient training data and background knowledge. I refer to these frequently mentioned entities as the *linguistic head*. Besides being frequent news topics, the mentions of head entities are also frequent (due to the volume of news) and the mention-to-entity dominance is very high.

However, at the same time there is a vast amount of *long-tail entities*, each different and with low frequency, that usually remain hard to resolve for any contemporary system. Support for this claim can be found in the related task of Word Sense Disambiguation. Here, the system accuracy on the most frequent word interpretations is close to human performance, while the least frequent words can be disambiguated correctly in at most 1 out of 5 cases (Postma et al., 2016a). It is my conviction that the linguistic long tail can never be fully tackled with further algorithmic inventions because these long-tail instances appear only incidentally and change continuously, and it is unlikely there will ever be sufficient training data. Even if we would increase the training data, it is impossible to guess the a priori distribution that applies to any actual test set across all the options (Postma et al., 2016a).

Additionally, probabilistic approaches do not employ any mechanisms to exclude anomalous interpretations. This leads to an explosion of potential interpretations which are dominated by the most popular ones even though they often

12 Hundreds of entity instances in DBpedia are named 'John Smith': `http://dbpedia.org/resource/John_Smith`.

13 `http://gerbil.aksw.org/gerbil/overview`

do not make any sense. In the NewsReader project,[14] for example, the most popular detected entity in 2.3 million news articles from 2003-2015 about the car industry was *Abraham Lincoln*, demonstrating how dominance leads to wrong and impossible interpretations (Vossen et al., 2016). This problem becomes even more substantial when we switch from the popular world represented in DBpedia to resolving long-tail entities, where the surface form ambiguity becomes too big to handle.[15] For instance, while *Ronaldo* can refer to only a few popular entities according to DBpedia, the number of people in the world that (have) share(d) this name is many orders of magnitude greater. For current systems, it is extremely hard to deal with this reality, while humans have no problem understanding news mentioning some non-famous *Ronaldo*. As these long-tail instances are only relevant within a specific context (time, location, topic, community), we need contextual knowledge and reasoning in order to decide which make sense.[16]

Types of knowledge In (MacLachlan and Reid, 1994), four types of contextual knowledge are defined that are essential for humans to interpret text. These are: *intratextual*, *intertextual*, *extratextual*, and *circumtextual* knowledge. Here, I relate these four categories to the EL task.

1. *Intratextual knowledge* is any knowledge extracted from the text of a document, concerning entity mentions, other word types (e.g., nouns, verbs), and their order and structure in the document. It relates to framing new and given information and notions such as topic and focus. Central entities in the discourse are referred to differently than peripheral ones.

2. *Extratextual knowledge* concerns any entity-oriented knowledge, found outside the document in (un)structured knowledge bases. Extratextual knowledge can be episodic (instantial) or conceptual. The former is the knowledge about a concrete entity: its labels, relation to other entities, and other facts or experiences. Conceptual knowledge refers to the expectations and knowledge gaps that are filled by an abstract model (i.e., ontology), representing relations between types of entities.

3. *Circumtextual knowledge* refers to the circumstances through which text as an artifact has come into existence. Documents are published at a specific time and location, written by a specific author, released by a certain publisher, and potentially belong to some series. These circumstances frame the written text and aid the interpretation of the mentioned entities.

4. *Intertextual knowledge* - Documents are not self-contained and rely on intertextual (cross-document) knowledge distilled by the reader from related

14 `http://newsreader-project.eu`

15 Probabilistic methods are sensitive to even small changes in the background knowledge: only switching to a more recent Wikipedia version causes a drop in performance because of the increased ambiguity and the change in knowledge distribution (Nguyen et al., 2016a).

16 These differ from the domain-specific entities (e.g., names of drugs in the medical domain), which are defined through a single contextual dimension (of topic) and do not necessarily suffer from knowledge scarcity.

> documents. They are published in a stream of information and news, assuming knowledge about preceding related documents, which typically share the same topic and community, and may be published around the same time and location. Early documents that introduce a topic typically make more explicit reference than those published later on when both the event and the topic have evolved.[17]

In this thesis, I will use the term *background knowledge* to refer to the union of the intertextual, circumtextual, and extratextual knowledge.

Many instances of intratextual knowledge are present in EL systems: information about surrounding words (word clouds) (Daiber et al., 2013; Yosef et al., 2011), entity order (Ilievski et al., 2016a), entity coreference (Ling et al., 2015a), substrings (Usbeck et al., 2014), abbreviations (Ilievski et al., 2016a), word senses (Moro et al., 2014), word relations (Cheng and Roth, 2013). Systems also tend to consider some extratextual knowledge, such as: entity-to-entity links (Daiber et al., 2013; Ilievski et al., 2016a; Ling et al., 2015a; Moro et al., 2014; Piccinno and Ferragina, 2014; Usbeck et al., 2014), entity labels (Daiber et al., 2013; Ilievski et al., 2016a; Ling et al., 2015a; Moro et al., 2014; Piccinno and Ferragina, 2014; Usbeck et al., 2014), semantic types (Ilievski et al., 2016a; Ling et al., 2015a), and textual descriptions (Daiber et al., 2013; Yosef et al., 2011). Considering the richness of extratextual knowledge found in public knowledge bases like DBpedia and Wikidata, its potential for reasoning seems much larger than what is currently exploited. For instance, one can build models over the instance-level knowledge in these KBs to capture implicit knowledge, which can be applied to enhance the scarce information that is directly available about a certain long-tail entity. To the best of my knowledge, the other two types of knowledge, circumtextual and intertextual knowledge, are systematically neglected in current systems. As such, it is no surprise that these systems fail to handle a case such as the *Hobbs murder* presented next.

Entity Linking in the long tail To illustrate my point, let us consider the following case. In the local news article titled "Hobbs man arrested in connection to nephew's murder",[18] a murder is reported that happened in Hobbs, New Mexico in 2016. It involves two long-tail entities: the killer Michael Johnson and its victim Zachariah Fields. Both entities have no representation in DBpedia, as they are not famous outside the context of this murder.

Current EL systems perform poorly on this document. For instance, Babelfy (Moro et al., 2014) links "Michael Johnson" to a retired American sprinter, "Johnson" to an American president, and "Zachariah" to a long-deceased religious clergyman and author from the 19th century. Not only are these interpretations incorrect, they are also highly incoherent from a human perspective: a retired sprinter, a 19th century religious author, and an ex-president are all identified in an article reporting a local murder in New Mexico in 2016.

What makes these interpretations silly to humans, but optimal to EL systems, is the different notion of coherence. Roughly, entity linkers define coherence via

17 Compare the use of hashtags in Twitter streams once an event becomes trending.

18 `https://goo.gl/Gms7IQ` Last visited: 18 April 2017

a probabilistic optimization over entity and word associations, resulting in interpretations that neither share context among themselves, nor with the document. Unlike machines, people employ rigorous contextual reasoning over time, location, topic, and other circumtextual knowledge about the article. Time would help to decide against the 19th century author as a victim in 2016. Similarly for location and topic: none of the system interpretations is related to Hobbs, New Mexico, or to any violent event. As systems do not use circumtextual knowledge, they have no human-like mechanisms to decide on improbable interpretations.

In addition, this document is not self-contained; it provides an update regarding an event that happened earlier and was already reported on. In such cases, its interpretation might benefit from (or even depend on) focused machine reading of earlier documents covering this topic. This is very natural for humans; still, current systems lack ways to obtain and integrate intertextual knowledge.

Evaluation The expected difficulties for systems to perform well on the distributional tail trigger the question to which extent is this tail captured in current evaluation datasets. It is unclear whether the present manner of evaluating entity identity is representative for the complex contextual dependency of the human language discussed in section 1.1. Considering that the evaluation datasets have been created by sampling human communication which is dominated by a small set of very frequent observations, my hypothesis is that the current evaluation exhibits a similar frequency bias. This bias towards the frequent and popular part of the world (*the head*) would be largely responsible for the performance of our EL tools. Presumably, EL tools are optimized to capture the head phenomena in text without considering the contextual constraints which are essential in order to deal with the less frequent phenomena in *the tail*. Understanding the evaluation bias better and improving its representativeness to challenge systems to perform well on the tail hence becomes a very important piece of the long-tail entities puzzle.

The focus of this thesis is on long-tail entities mentioned in news documents. Notably, long-tail entities are also problematic in other text modalities, particularly in social media data (Derczynski et al., 2016), as well as in other entity-centric tasks, e.g., mapping queries to the Linked Open Data cloud (Meij et al., 2011).

1.3 RESEARCH QUESTIONS

As we have seen, establishing identity is a relevant and timely topic, but also one that is multifaceted and extremely challenging. For a small portion of entities (*the frequent head*), there are vast amounts of knowledge in both structured and unstructured form. At the same time, the relevance of most entities is restricted to a specific context, i.e., a specific combination of topic, time, space, and community. These entities constitute *the rare long tail*: they have no representation in encyclopedias, there is little to none news reporting on them, and distinguishing them from other contextual entities might require more effort than in the case of the famous entities. Considering the discussion on language pragmatics and

background knowledge earlier in this chapter, I expect that current NLP technology especially struggles to establish identity of long-tail entities. The research question that is investigated in this thesis is:

> *How can the performance of NLP techniques to establish identity of long-tail cases be improved through the use of background knowledge?*

I address the main research question by addressing five aspects of the task of establishing identity of long-tail entities: 1. description and observation 2. analysis of the evaluation bias 3. improvement of the evaluation bias 4. access to knowledge 5. role of knowledge. These are transcribed into the following five subquestions that will be addressed within the chapters of this thesis:

RQ1 (Description and observation): *How can the tail entities be distinguished from head entities?*

RQ2 (Analysis of the evaluation bias): *Are the current evaluation datasets and metrics representative for the long-tail cases?*

RQ3 (Improvement of the evaluation bias): *How can we improve the evaluation on the long-tail cases?*

RQ4 (Access to knowledge): *How can the knowledge on long-tail entities be accessed and enriched beyond DBpedia?*

RQ5 (Role of knowledge): *What is the added value of background knowledge models when establishing the identity of long-tail entities?*

In the next section, I provide an overview of the approach taken in this thesis to address these five research questions.

1.4 APPROACH AND STRUCTURE OF THE THESIS

To understand the task of establishing identity of entities in text and the nature of its difficulty, I make an effort to categorize all entities into one of the following three categories:

1. **Head entities (E1)** are frequent interpretations associated with a certain entity mention. For example, the form 'Paris' is typically used to refer to the French capital Paris, which means that the instance `http://dbpedia.org/page/Paris` belongs to the distributional head. In practice, these entities have a representation in DBpedia and can be expected to be disambiguated accurately by existing EL tools.

2. **Tail entities (E2)** are infrequent interpretations of an entity mention and rely on a richer context in order to be disambiguated correctly. For instance, the city of Paris in Texas can be referred by the form 'Paris', but this does not occur as common in communication as with the French capital with

the same name. Hence, `http://dbpedia.org/page/Paris,_Texas` is considered a tail interpretation. E2 entities can be operationalized through those entities that have a representation in DBpedia, but pose a greater challenge to existing tools compared to E1 entities.

3. **NIL entities (E3)** are extremely infrequent in communication and are relevant only within concrete contextual circumstances.[19] An example for an E3 entity is a local bar in Amsterdam called 'Paris'. Most entities in the world fall into this category. Due to their numerousness and restricted relevance, E3 entities cannot be expected to have an existing representation in DBpedia, or even broader in the Linked Open Data cloud. The available knowledge about these entities is non-redundant and incomplete, whereas their ambiguity is potentially very high, which intuitively makes them difficult to resolve by existing EL tools.

The term *long-tail entities* will be used throughout this thesis to refer to E2 and E3 entities simultaneously. Given that both DBpedia and the evaluation datasets can be seen as proxies of human communication, it can be expected that the well-represented entities in DBpedia (E1) are most prominent in evaluation datasets, whereas the long-tail entities (of type E2 and E3) are rarely evaluated. At the same time, E2 and E3 entities are extremely important, because current systems underperform on them. I argue that the challenge in resolving long-tail entities is a consequence of several factors:

1. datasets do not cover these cases sufficiently, leading to semantic overfitting;
2. systems have no access to the needed instance-level knowledge on these cases, when they have a representation (E2);
3. systems apply shallow threshold-based strategies to deal with E3 entities, instead of rich and diverse human-like knowledge.

1.4.1 *Describing and observing the head and the tail*

Chapter 2 addresses *RQ1* and covers the entities of type E1 and E2. Previous research has often hypothesized that tail cases are much harder and much less frequent than head cases, without specifying what each of these entails. In this chapter, we fill this gap by describing the head and the tail of the entity linking task.[20] We expect that certain entities (tail) are more challenging to establish than others (head) and we seek to define the dimensions along which this distinction can be made. These hypotheses are then to be tested against current state-of-the-art technology to assess their validity.

19 Our definition of E3 entities corresponds to the notion of NIL entities (Ji and Grishman, 2011). I will use these two terms interchangeably in this thesis.

20 Note: Given that most of the research reported in this thesis could not possibly have existed without the contribution of my supervisors and colleagues, in the remainder of this thesis I will regularly talk about 'we' to signify the joint effort of me and my paper co-authors.

The content of this chapter is based on research published in the following publication:

1. Filip Ilievski, Piek Vossen, and Stefan Schlobach (2018). "Systematic Study of Long Tail Phenomena in Entity Linking." In: *The 27th International Conference on Computational Linguistics (COLING 2018)*

1.4.2 *Analyzing the evaluation bias on the long tail*

Chapter 3 addresses *RQ2* and focuses on the entities of type E2 and E3. Here we analyze whether the current evaluation datasets and metrics are representative for the long-tail cases. In this chapter, we pick five semantic NLP tasks of disambiguation and reference (including Entity Linking), and apply a set of model-based metrics to assess the representativeness of their evaluation datasets.

The content of this chapter is based on research published in the following two publications:

1. Filip Ilievski, Marten Postma, and Piek Vossen (2016c). "Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text?" In: *The 26th International Conference on Computational Linguistics (COLING 2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1180–1191. URL: http://aclweb.org/anthology/C16-1112

2. Marieke Van Erp, Pablo N Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis (2016). "Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Vol. 5, p. 2016

1.4.3 *Improving the evaluation bias on the long tail*

Chapter 4 aims to improve the evaluation on the long-tail cases (E2 and E3), thus addressing *RQ3*. Once the distinction between 'easy' head cases and 'difficult' tail cases has been made, and we have analyzed the representativeness bias of current evaluation datasets, we move forward to create the first task that deliberately focuses on long-tail identity, and we analyze its impact.

The content of this chapter is based on the research published in the following four publications:

1. Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp (2016b). "Moving away from semantic overfitting in disambiguation datasets." In: *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. Austin, TX: Association for Computational Linguistics, pp. 17–21. URL: http://aclweb.org/anthology/W16-6004

2. Marten Postma, Filip Ilievski, and Piek Vossen (2018). "SemEval-2018 Task 5: Counting Events and Participants in the Long Tail." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics

3. Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers (2018a). "Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*

4. Piek Vossen, Marten Postma, and Filip Ilievski (2018b). "ReferenceNet: a semantic-pragmatic network for capturing reference relations." In: *Global Wordnet Conference 2018, Singapore*

1.4.4 *Enabling access to knowledge about long-tail entities beyond DBpedia*

Chapter 5 deals with the limited access to knowledge about E2 and E3 entities. Therefore, this chapter addresses *RQ4*. As discussed above, long-tail entities are characterized by scarce to no accessible knowledge at all. At the same time, a field that is centered around the notion of identity is Semantic Web, offering billions of facts about various entities. We investigate how this knowledge can be made accessible for consumption by NLP tools and make an attempt to improve its accessibility.

The content of this chapter is based on the research published in the following four publications:

1. Filip Ilievski, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2015). "LOTUS: Linked Open Text UnleaShed." In: *Proceedings of the Consuming Linked Data (COLD) workshop*

2. Filip Ilievski, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2016b). "LOTUS: Adaptive Text Search for Big Linked Data." In: *European Semantic Web Conference (ESWC) 2016*. Springer International Publishing, pp. 470–485

3. Wouter Beek, Laurens Rietveld, Filip Ilievski, and Stefan Schlobach (2017). "LOD Lab: Scalable Linked Data Processing." In: *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering*. Lecture notes of summer school. Springer International Publishing, pp. 124–155

4. Wouter Beek, Filip Ilievski, Jeremy Debattista, Stefan Schlobach, and Jan Wielemaker (2018). "Literally better: Analyzing and improving the quality of literals." In: *Semantic Web Journal (SWJ)* 9.1, pp. 131–150. DOI: 10.3233/SW-170288. URL: https://doi.org/10.3233/SW-170288

1.4.5 *The role of knowledge in establishing identity of long-tail entities*

Chapter 6 addresses *RQ5* and it concerns the entities of type E3. Most entities have no accessible representation, which prevents one to establish their identity. We argue that the extreme ambiguity representing the long tail can only be addressed by robust reasoning over well-targeted, but rich, contextual knowledge. Special attention should be devoted to filling gaps in background knowledge by implicit values. We investigate the idea of establishing identity among unknown entities from text by enhancing the explicit information given in text with background knowledge models which capture implicit expectations.

The content of this chapter is a combination of new research and the content published in the following two publications:

1. Filip Ilievski, Piek Vossen, and Marieke van Erp (2017). "Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking." In: *International Conference on Language, Data and Knowledge*. Springer, Cham, pp. 143–149
2. Filip Ilievski, Eduard Hovy, Qizhe Xie, and Piek Vossen (2018). "The Profiling Machine: Active Generalization over Knowledge." In: *ArXiv e-prints*. arXiv: `1810.00782 [cs.AI]`

1.5 SUMMARY OF FINDINGS

The main findings of this thesis, providing answers to the research questions RQ1-RQ5, are as follows:

F1. There is a positive dependency of system performance on frequency and popularity of instances, and a negative one with ambiguity of surface forms. Essentially, this confirms the intuition that system performance is largely based on head cases, and declines strongly on the tail.

F2. The commonly used datasets to evaluate disambiguation and reference NLP tasks suffer from low ambiguity, low variance, high dominance, and limited temporal spread.

F3. On a deliberately created task to evaluate tail instances, we observe very low accuracy of the participating systems. This shows that dealing with high ambiguity and not being able to rely on frequency biases, poses a great challenge for current NLP systems.

F4. By creating LOTUS, we provide the Linked Open Data community with the largest centralized text index and access point to the LOD Laundromat data collection. This allows EL systems to use the knowledge found among the billions of statements of the LOD Laundromat collection, thus essentially increasing their recall on the tail instances.

F5. Neural background knowledge models ("profiling machines") are built and applied in order to complete the partial profiles extracted from text and establish their identity. The evaluation of these machines on the task of establishing long-tail identity in text shows promising results when applied on top of automatically extracted information from text. We observe no clear patterns between the effectiveness of our profilers and the data ambiguity.

1.6 SOFTWARE AND DATA

- https://github.com/cltl/Profiling Code from the paper 'The Profiling Machine: Active Generalization over Knowledge' , where we build two neural architectures for generating expectations for implicit/missing values of people profiles.
- https://github.com/cltl/AnnotatingEntityProfiles All code from the annotation tool that was used to mark evidence for entity properties in text, but also complement the structured data with new knowledge found in text.
- https://github.com/cltl/LongTailAnnotation The entire code from the first version of this tool, used to annotate mentions of events in text based on structured knowledge.
- https://github.com/cltl/EL-long-tail-phenomena Analysis of various long-tail properties of entity linking datasets in relation to system performance.
- https://github.com/filipdbrsk/LOTUS_Indexer All code that was used to create the LOTUS index, the largest public text search index to the LOD cloud.
- https://github.com/filipdbrsk/LOTUS_Search All code that powers the search API and website of LOTUS (the largest public text search index to the LOD cloud), once the index is already created.
- https://github.com/D2KLab/relink Entity Linking system based on several human-inspired heuristics for domain adaptation, written in Python.
- https://github.com/cltl/HumanLikeEL Entity linking system inspired by human classification of knowledge types for framing.
- https://github.com/cltl/News2RDF A pipeline for conversion of news documents to RDF knowledge.
- https://github.com/cltl/SemEval2018-5_Postprocessing Postprocessing scripts for SemEval-2018 task 5.
- https://github.com/cltl/GunViolenceCorpus The Gun Violence Corpus, mostly annotated during the organization of the SemEval-2018 task 5.

- https://github.com/cltl/LongTailQATask The entire code for the SemEval-2018 task 5.
- https://github.com/cltl/SemanticOverfitting The code used to analyze datasets from five disambiguation and reference tasks using a set of statistics.
- https://github.com/globalwordnet/ili RDF backend of the Wordnet Inter-Lingual Index.
- https://github.com/dbpedia-spotlight/evaluation-datasets Analysis of evaluation datasets for Entity Linking.
- https://github.com/filipdbrsk/LODLanguages Analysis of the state of the language literals in the LOD cloud.
- https://github.com/cltl/python-for-text-analysis Material of the course 'Python for text analysis'.

2 DESCRIBING AND OBSERVING THE HEAD AND THE TAIL OF ENTITY LINKING

As discussed in chapter 1, it is customary to establish identity of entities by linking them to an existing semantic representation, formally known as the task of entity linking. The task of Entity Linking (EL) anchors recognized entity mentions in text to their semantic representation, thus establishing identity and facilitating the exploitation of background knowledge, easy integration, and comparison and reuse of systems. EL typically makes a distinction between two types of mentions: linkable mentions, for which there is a corresponding representation in a referent knowledge base, and NILs, whose representation is not (yet) found in a knowledge base. We will focus on NIL (E3) entities in chapter 6. Here, we invest an effort to understand whether some (classes of) linkable entities are more difficult to link than others, and if so, how can these two groups (named E1 and E2 in chapter 1) be formally described. This chapter thus addresses RQ1, the first research question we put forward in this thesis: *How can the tail entities (E2) be distinguished from head entities (E1)?*

The content of this chapter is based on research published in the following publication:

1. Filip Ilievski, Piek Vossen, and Stefan Schlobach (2018). "Systematic Study of Long Tail Phenomena in Entity Linking." In: *The 27th International Conference on Computational Linguistics (COLING 2018)* In this paper I was the main contributor, posed the problem, designed the analysis, and implemented it.

2.1 INTRODUCTION

The past years featured a plethora of EL systems: DBpedia Spotlight (Daiber et al., 2013), WAT (Piccinno and Ferragina, 2014), AGDISTIS (Moussallem et al., 2017), to name a few. These systems propose various probabilistic algorithms for graph optimization or machine learning, in order to perform disambiguation, i.e., to pick the correct entity candidate for a surface form in a given context. The reported accuracy scores are fairly high, which gives an impression that the task of EL is both well-understood and fairly solved by existing systems.

At the same time, several papers (Esquivel et al., 2017; Ilievski et al., 2016c, 2017; Van Erp et al., 2016) have argued that state-of-the-art EL systems base their performance on frequent 'head' cases, while performance drops significantly when moving towards the rare 'long tail' entities. This statement seems intuitively obvious, but no previous work has quantified what the 'head' and 'tail' of EL entails. In fact, the lack of definition of head and tail in this task

prevents the (in)validation of the hypothesis that interpreting some (classes of) cases is more challenging for systems than others. This, in turn, means that we are currently unable to identify the difficult cases of EL for which current systems need to be adapted, or new approaches need to be developed. Previous linguistic studies which analyze word distributions (Corral et al., 2015; Kanwal et al., 2017; Zipf, 1935) cannot be applied for this purpose, because they do not study EL, nor the relation of the head-tail distribution to system performance.

Understanding the long-tail cases better and explicitly addressing them in the design of EL systems will be beneficial because: 1. long-tail entities are common in textual data, and are thus relevant for various use cases. 2. unlike the head entities, the knowledge about the tail entities is less accessible (in structured or unstructured form), not redundant, and hard to obtain. 3. to perform well on the tail, systems are required to interpret entity references without relying on statistical priors, but by focusing on high-precision reasoning.

This chapter addresses the question: **how can the tail entities (E2) be distinguished from head entities (E1)?** Namely, we are interested which data properties capture the distinction between the head and the tail in entity linking, and to what extent. The main contributions of this chapter are the following:

1. We define the head-tail properties of entity linking **(Section 2.3)**.[1] This is the first work that looks systematically into the relation of surface forms in evaluation datasets and their instances in DBpedia, and provides predictions in the form of a series of hypotheses about head-tail phenomena.

2. We analyze existing entity linking datasets with respect to these head-tail properties, demonstrating that data properties have certain correlations that follow our hypotheses **(Section 2.4)**.

3. We describe how the performance of systems correlates with head and tail cases, proving that the head-tail phenomena and their interactions influence the performance of systems **(Section 2.5)**.

4. We provide recommendations on how to address the tail in future EL research (**Section 2.7**).

2.2 RELATED WORK

In the related task of Word Sense Disambiguation (WSD), Postma et al. (2016a) analyzed the impact of frequency on system accuracy, showing that the accuracy on the most frequent words is close to human performance (above 80%), while the least frequent words can be disambiguated correctly in at most 20% of cases. We note that no past work has performed such analysis for the task of entity linking. In addition, while the sense inventory used for WSD, WordNet, is known to be limited in coverage (Jurgens and Pilehvar, 2016), it is even harder to create a

1 We consider the following properties: ambiguity of surface forms, variance of instances, frequency of surface forms, frequency of instances, and popularity of instances.

comprehensive, corpus-independent inventory for the EL task, given its extreme real-world ambiguity.[2]

As we will see in chapter 3, the well-known datasets for semantic NLP tasks exhibit very low ambiguity and variation, and a notable bias towards dominance. Overall, tasks and datasets show strong semantic overfitting to the head of the distribution (the popular part) and are not representative for the diversity of the long tail. Specifically, we show that EL datasets contain very little referential ambiguity and evaluation is focused on well-known entities, i.e., entities with high PageRank (Page et al., 1999) values.

Ji and Grishman (2011) define the entities that have no representation in a referent knowledge base as **NIL entities**. These correspond to our E3 entity category in chapter 1. E3/NIL entities are typically considered to have low frequencies within a corpus and/or to be domain-specific. While NILs are a challenge that concerns the long tail of EL, in this chapter we focus on those entities that have been linked to Wikipedia, but are still infrequent, since this provides an entry for extensive analysis of their properties. We return to the NIL entities in chapter 6.

Van Erp et al. (2015) described **dark entities** as entities for which no relevant information is present in a given knowledge base, thus expanding the notion of NIL entities to also include cases where an entity representation exists, but it is insufficient to reason over. From a perspective of our entity categorization in chapter 1, the dark entities can be seen as a joint term for E2 and E3 entities.

We distinguish emerging and domain entities. **Emerging entities** are time-bound entities, recently unknown but potentially becoming popular in news in a short time (Graus et al., 2018; Hoffart et al., 2014). A body of work has dealt with **domain entities**, whose relevance is restricted within a topic, e.g., biomedical (Zheng et al., 2015), or historical domain (Heino et al., 2017). While emerging and domain entities mostly make up the tail in EL, their definition is orthogonal to our work. We strive to distinguish the head and the tail in EL based on data properties that capture the interaction between surface forms and their denoted instances (ambiguity, variance, frequency, and popularity; explained in the next section), and avoid a discussion on the distinction between head or tail in a categorical way.

Finally, studying distributional properties of entity expressions and their linking, as we do in this study, is different from the classical linguistic studies on the distribution of words (Corral et al., 2015; Kanwal et al., 2017; Zipf, 1935). Linked entity data provides information on the surface forms, the meaning, and the referent of the surface form, whereas distributional studies on words only provide knowledge on the surface forms and to a limited extent on their sense, but never on their reference.

2 As an illustration, the amount of possible meanings of an ambiguous word, like play, is in the range of dozens or potentially hundreds. The amount of possible referents of "John Smith" is orders of magnitude larger and even difficult to estimate.

> " Washington announces Alex Smith trade
> It seems like months ago that the Chiefs traded Alex Smith to Washington .
> ...
> Smith , 33, originally entered the league as the No. 1 overall choice in 2005. The 49ers traded him to the Chiefs on March 12, 2013. Now, he moves again."

Figure 1: Snippet from a news article describing an event in the American Football League, NFL. The highlighted phrases mark entity surface forms. Each color denotes a separate instance. Source: `https://profootballtalk.nbcsports.com/2018/03/14/washington-announces-alex-smith-trade/`

2.3 APPROACH

To address our research goal of **distinguishing the head and the tail of EL**, we first explain the notions of ambiguity, variance, frequency, and popularity. Next, we formulate a set of hypotheses regarding their interaction and our expected influence on system performance. We also describe our choice of data collections and EL systems to analyze. The code of this analysis is available on GitHub at `https://github.com/cltl/EL-long-tail-phenomena`.

2.3.1 *The head-tail phenomena of the entity linking task*

Each entity exists only once in the physical world. However, this is different in our communication where: 1. certain instances are very prominent and others are mentioned incidentally, resulting in a distribution over mentions of instances; 2. certain surface forms are very prominent and others occur only rarely, resulting in a distribution over surface forms to make reference. The task of EL covers a many-to-many relation between surface forms observed in text and their instances potentially found in a knowledge base. Surface forms and instances both have their own **frequency distribution**, pointing to the same underlying Grice (1975) mechanisms, governed by an envisioned trade-off between efficiency and effectiveness. The Gricean maxims of quantity and relation are especially adequate to explain this phenomenon. The maxim of quantity explains that the speaker/writer would transmit sufficient amount of information, but not more than that. The maxim of relation dictates that the speaker/writer transmits information that is relevant and pertinent to the discussion.

Considering the frequency distributions of surface forms and instances, as well as their complex relation governed by pragmatic principles, we can define the following phenomena along which we try to distinguish head and tail cases in EL: frequency of surface forms, ambiguity of surface forms, frequency of in-

stances, popularity of instances, and variance of instances. We illustrate these phenomena on the text snippet shown in Figure 1, which describes an event about the American Football League, NFL.

FREQUENCY OF SURFACE FORMS In our communication, surface forms have various frequency of occurrence. As we have no access to the overall human communication, we approximate this with available textual documents. Thus, the frequency of a surface form is equal to its number of occurences in a corpus of textual documents. In Figure 1, the forms "Washington", "Chiefs", and "Alex Smith" all have a frequency of 2, while the frequency of "Smith" and "49ers" is 1.

Frequent surface forms include "U.S." and "Germany", but also "John Smith". The frequency of a surface form can be explained by its relation to one (or a few) very frequent/popular instances (`United States`), but it can also be a result of high ambiguity ("John Smith" is a common name, so it simply refers to many possible instances).

AMBIGUITY OF SURFACE FORMS The notion of ambiguity captures the amount of distinct instances that a surface form refers to. Since we have no access to the true ambiguity for a surface form, we approximate it by counting its instances in a corpus of textual documents - we call this **observed ambiguity (OA)**. EL requires disambiguation with respect to a resource, hence the upper bound for the OA of a form is the ambiguity of that form in a resource (typically DBpedia) - we call this **resource ambiguity**. In this chapter we analyze observed ambiguity, and we return to the resource ambiguity in chapter 3.

Non-ambiguous forms are the ones that have been observed to refer to a single instance in an entire text collection. Ambiguous forms, then, are those that have been observed to refer to at least two different instances. In the snippet in Figure 1, all forms are non-ambiguous, i.e., they have an ambiguity of 1. Provided that, for example, the second mention of "Washington" would have referred to the city of Washington D.C., the ambiguity of this form would have been 2.

FREQUENCY OF INSTANCES Similarly to the frequency of forms, the frequency of instances can be counted in textual datasets, as proxies for referential communication. The frequency of an instance refers to its number of occurrences in a corpus. In our example, the instance `Alex_Smith` has a frequency of 3, whereas
`Washington_Redskins` and `Kansas_City_Chiefs` have a frequency of two, and `San_Francisco_49ers` has a frequency of 1.

POPULARITY OF INSTANCES The instance frequency can be seen as a way of capturing popularity of instances through textual information. A more appropriate way to quantify popularity might be through the volume

of knowledge about an instance found in a knowledge base. We consider DBpedia to be representative for the general-purpose world knowledge. Concretely, we make use of PageRank values for DBpedia entities, computed over the DBpedia knowledge graph following the algorithm in (Thalhammer and Rettinger, 2016).[3] For instance, the popularity of `Washington_Redskins`, measured with PageRank, is 89.97, whereas the popularity of `Alex_Smith` is 4.26.

For an instance, we expect its frequency in a corpus and its popularity in a knowledge base to be related, as the instances which are more popular would be on average mentioned more frequently than others. In addition, frequent/popular instances tend to participate in metonymy relations with other instances topically related to them. For instance, `United_States` as a country relates to `United_States_Army` and to `Federal_government_of_the_United_States` - two other entities of a different type, but possibly referenced by similar surface forms.

VARIANCE OF INSTANCES The variance of an instance is measured by the number of distinct forms that refer to it. Like ambiguity, we consider the variance as observed in a textual dataset, which we call **observed variance**. The observed variance, however, can in theory be higher than the variance as found in a resource (**resource variance**). Here we focus on the observed variance; we measure the resource variance in chapter 3.

We expect that frequent and popular instances exhibit high variance, as these are intuitively quite prominent and relevant, very often across many different circumstances, and are typically referred to by a relatively wide set of surface forms.

In Figure 1, `Alex_Smith` is referred by two forms ("Alex Smith" and "Smith"), leading to a variance of two. All other instances have a variance of 1.

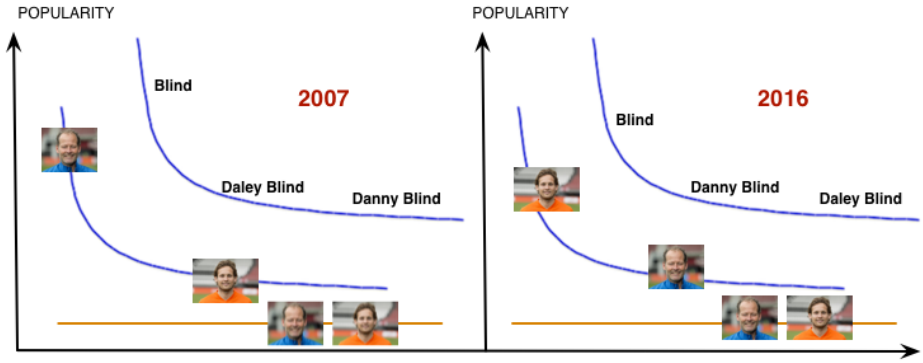
To illustrate the complexity and relevance of these head-tail properties in our reporting about the world, let us consider two Dutch soccer players: Danny Blind[4] and his son, Daley Blind[5]. Danny Blind retired from professional football in 1999, whereas his son had his debut in professional soccer in 2008. Figure 2 presents three distributions of these two entity instances: their physical distribution in the world (orange line), their popularity distribution (lower blue line), and the frequency of the surface forms that refer to them (upper blue line). On a physical level, we see no difference between any two entities: each exists only once in the physical world. The distributions of surface forms and instances in communication, however, are skewed towards a small number of very frequent head cases, and contain a large tail of infrequent cases.

3 The precomputed PageRank values can be downloaded from `http://people.aifb.kit.edu/ath/`. They can also be queried through the official DBpedia SPARQL endpoint: `http://dbpedia.org/sparql`.

4 `http://dbpedia.org/resource/Danny_Blind`

5 `http://dbpedia.org/resource/Daley_Blind`

Figure 2: Distribution of forms and instances relating to people with a surname 'Blind', in particular the soccer players Daley and Danny Blind. Three distributions of these two entity instances are shown: their physical distribution in the world (orange line), their popularity distribution (lower blue line), and the frequency of the surface forms that refer to them (upper blue line). The entities are ranked (left to right) based on their popularity at the specific year.



For forms that are dominantly used to refer to a popular instance, the frequency of instances and the frequency of forms coincide to some extent. However, the aspects of ambiguity and variance 'weaken' this relation, as other forms can also refer to the same instance, and the same (or a similar) form can be used to refer to a large number of instances. This relation is additionally dependent on the spatial, temporal, and topical context within which this interaction between forms and instances is considered. As Figure 2 shows, the instance distribution can easily change over time, as some entities become more and others less relevant over time. For instance, prior to the professional debut of Daley Blind, the form 'Blind' would typically refer to his father, Danny; whereas nowadays this has shifted because the current relevance of Daley Blind exceeds his father's. The change in the instance distribution does not necessarily need to cause a change in the distribution of the surface forms, and vice versa. Similar claims can be made about other contextual dimensions, like topic or space. We look further into the contextual quantification of datasets in chapter 3.

Next, we put forward a series of pragmatics-governed hypotheses that describe the head-tail phenomena and their interaction.

2.3.2 *Hypotheses on the head-tail phenomena of the entity linking task*

We look systematically at the relation of surface forms in datasets and their instances in DBpedia, and provide a series of hypotheses regarding the head-tail phenomena and their relation to system performance (Table 1). Some of these hypotheses, e.g., D1 and D2, are widely accepted as common knowledge but

ID	Hypothesis	Sec
D1	Only a few forms and a few instances are very frequent in corpora, while most appear only incidentally.	2.4.1
D2	A few instances in corpora are much more popular (have much higher PageRank) compared to most other.	2.4.2
D3	Only a small portion of all forms in corpora are ambiguous, i.e., the observed ambiguity is low.	2.4.3
D4	Only a small portion of all instances in corpora are referred to with multiple forms, i.e., the observed variance is low.	2.4.4
D5	There is a positive correlation between observed ambiguity of forms and their frequency.	2.4.5
D6	There is a positive correlation between observed variance of instances and their frequency.	2.4.5
D7	There is a positive correlation between observed variance of instances and their popularity.	2.4.5
D8	There is a positive correlation between popularity of instances and their frequency.	2.4.5
D9	The frequency distribution within all forms that refer to an instance is Zipfian.	2.4.6
D10	The frequency and the popularity distribution within all instances that a form refers to is Zipfian.	2.4.6
S1	Systems perform worse on forms that are ambiguous than overall.	2.5.1
S2	There is a positive correlation between system performance and frequency/popularity.	2.5.2
S3	Systems perform best on frequent, less ambiguous forms, and worst on infrequent, ambiguous forms.	2.5.3
S4	Systems perform better on ambiguous forms with imbalanced, compared to balanced, instance distribution.	2.5.4
S5	Systems perform better on frequent instances of ambiguous forms, compared to their infrequent instances.	2.5.4
S6	Systems perform better on popular instances of ambiguous forms, compared to their unpopular instances.	2.5.4

Table 1: Hypotheses on the data properties (D*) and on their relation to system performance (S*).

have rarely been investigated in EL datasets. Others, such as S4 and S5, are entirely new.

2.3.3 *Datasets and systems*

We focus on well-known EL datasets with news documents, preferring larger sets with open licenses. Many customary EL datasets are however quite small ($< 1,000$ mentions). We opted to perform our analysis on the following two data collections, with five corpora in total:

AIDA-YAGO2 (Hoffart et al., 2011) - we consider its train, test A, and test B sets, summing up to 34,929 entity forms in 1,393 news documents, published by Reuters from August 1996 to August 1997.

N3 (Röder et al., 2014) is a collection of three corpora released under a free license. We consider the two N3 corpora in English: RSS-500 and Reuters-128. Reuters-128 contains economic news published by Reuters, while RSS-500 contains data from RSS feeds, covering various topics such as business, science, and world news. These two corpora consist of 628 documents with 1,880 entity forms in total.

We analyzed the EL performance of recent public and open-sourced entity linkers, as the state-of-the-art:

AGDISTIS [6] (Moussallem et al., 2017)[7] combines graph algorithms with context-based retrieval over knowledge bases.

DBPEDIA SPOTLIGHT (Daiber et al., 2013)[8] is based upon cosine similarities and a modification of TF-IDF weights.

WAT (Piccinno and Ferragina, 2014)[9] combines a set of algorithms, including graph- and vote-based ones.[10]

2.3.4 *Evaluation*

We apply the customary metrics of precision, recall, and F1-score to measure system performance in Section 2.5. In Table 2, we briefly describe the computation of true positives (TPs), false positives (FPs), and false negatives (FNs) per entity instance class. For example, if the gold instance is C_1, then we count a TP when the system instance is also the instance C_1. In case the system instance is different, C_i, $i \neq 1$, this leads to a FN for C_1 and a FP for C_i. A special entity

6 The official name of this system is Multilingual AGDISTIS (MAG).

7 AGDISTIS API: `http://akswnc9.informatik.uni-leipzig.de:8113/AGDISTIS`, used on 24/05/2018.

8 Spotlight API: `http://spotlight.fii800.lod.labs.vu.nl/rest/disambiguate`, used on 24/05/2018.

9 WAT API: `https://wat.d4science.org/wat/tag/json`, used on 24/05/2018.

10 All three APIs link to the Wikipedia dump from April 2016. Since the official DBpedia Spotlight endpoint at `http://model.dbpedia-spotlight.org/en/disambiguate` links to a newer Wikipedia version (February 2018 at the time of writing of the underlying paper for this chapter), we set up our own endpoint that performs linking to the model 2016-04, to enable fair comparison with the other two systems. We reached similar conclusions with both versions of DBpedia Spotlight.

G\S	C_1	...	C_N	NILL
C_1	TP(1)		FP(N), FN(1)	FN(1)
...				
C_N	FP(1), FN(N)		TP(N)	FN(N)
NILL	FP(1)		FP(N)	-

Table 2: Computation of True Positives (TPs), False Positives (FPs), and False Negatives (FNs) per entity instance class $C_1, \ldots, C_N$. 'G'=gold instance, 'S'=system instance.

class are the NILs: predicting a NIL case incorrectly by the system results in a FN for the correct class; inversely, if the system was supposed to predict a NIL and it did not, then we count a FP. In our analysis we exclude the cases referring to NILs.

2.4 ANALYSIS OF DATA PROPERTIES

2.4.1 *Frequency distribution of forms and instances in datasets*

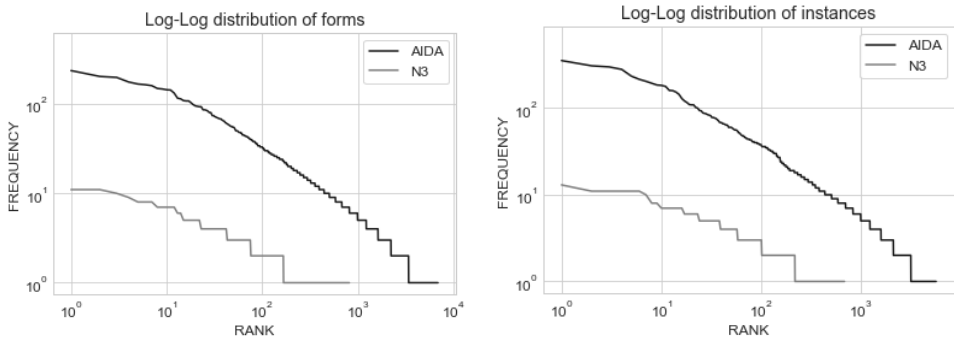
Zipf's law (Zipf, 1935) dictates that in a given textual dataset, the frequency of any word is inversely proportional to its rank in a frequency table. According to this law, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.: the rank-frequency distribution is an inverse relation.

Our hypothesis *D1* represents a variation of Zipf's law for the EL task: we expect that this rank-frequency relation holds also to a large extent for entity surface forms and for entity instances. In other words, we hypothesize that only a few forms and a few instances are very frequent in corpora, while most appear only incidentally.

The log-log frequency distributions of forms and instances (Figure 4a and 4b) show an almost ideal Zipfian distribution in the case of AIDA (with a slope coefficient of -0.9085 for forms and -0.9657 for instances) and to a lesser extent for N3 (a slope of -0.4291 for forms and -0.5419 for instances). The less Zipfian curves of N3 are probably because this data collection is significantly smaller than AIDA.

The similar shape of the form and the instance distribution per dataset can be explained by the dependency between these two aspects. Namely, the form 'U.S.' denoting the instance `United_States` 462 times is reflected in both the form and the instance distributions. However, these two distributions are only identical if the ambiguity and variance are both 1. In practice, the mapping between forms and instances is M-to-N, i.e., other forms also denote `United_States` (such as 'America') and there are other instances referred to by a form 'US' (such as `United_States_dollar`).

Figure 3: Log-log distribution of form frequency and instance frequency.



(a) form frequency

(b) instance frequency

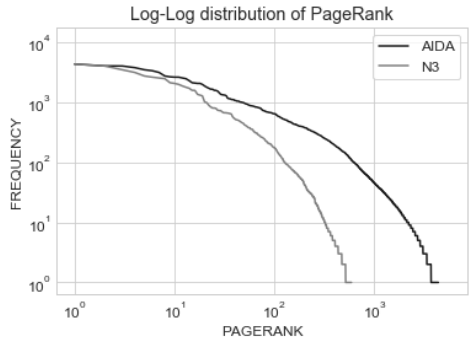


Figure 4: Log-log distribution of PageRank.

2.4.2 *PageRank distribution of instances in datasets*

Similar to the instance frequency, we expect that a few instances in the corpora have an extremely high PageRank compared to most others ($D2$). Figure 4 shows the PageRank distribution of our two datasets. We observe that most entity mentions in text refer to instances with a low PageRank value, while only a few cases have a high PageRank value. Not surprisingly, the instance with the highest PageRank value (`United_States`) is at the same time the instance with the highest corpus frequency.

We inspect the effect of frequency and PageRank on system performance in Section 2.5.2.

2.4.3 *Ambiguity distribution of forms*

We hypothesize that only a small portion of all forms in a corpus are ambiguous ($D3$). As shown in Table 3, when both datasets are merged and NIL entities

excluded, only 508 surface forms (6.73%) are ambiguous, as opposed to 7,037 monosemous forms (93.27%). This extremely high percentage validates our hypothesis. Moreover, in Sections 2.5.1, 2.5.3, and 2.5.4, we show that it also has a strong effect on system performance.

	1	2	3	4	5	6	..	12
AIDA	6,400	359	78	29	7	3		1
N3	794	18	2	1	0	0		0
BOTH	7,037	381	84	29	10	3		1

Table 3: Ambiguity distribution per dataset, after NILs are excluded. Columns represent degrees of ambiguity. Note that the values in the 'BOTH' row are not simply a sum of the ambiguity values of the individual datasets. This is because certain forms occur in both datasets, and the instances denoted by these forms are not necessarily identical across the two datasets.

2.4.4 *Variance distribution of instances*

We expect that only a small portion of all instances in a corpus are referred to with multiple forms (*D4*). The results of our variance analysis are given in Table 4. Over both datasets, 1,568 instances (25.61%) are referred to by multiple forms, as opposed to 4,555 instances (74.39%) which are always referred to by the same form. While the distribution of variance is much more even compared to that of ambiguity, we observe that most of the instances have a variance of 1.[11]

	1	2	3	4	5	6	7	8	9	10	11
AIDA	4,156	1,118	230	56	19	10	6	0	1	1	1
N3	550	106	15	7	1	0	0	0	0	0	0
BOTH	4,555	1,206	247	74	22	10	6	0	0	2	1

Table 4: Variance of instances with respect to the number of surface forms that reference them. Columns represent degrees of variance. Note that the values in the 'BOTH' row are not simply a sum of the variance values of the individual datasets. This is because certain instances occur in both datasets, and the forms used to refer to them are not necessarily identical in the two datasets.

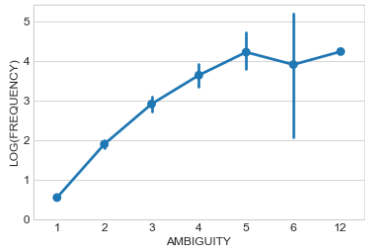
2.4.5 *Interaction between frequency, PageRank, and ambiguity/variance*

In the previous four Sections we analyzed the frequency distribution of individual data properties. Here we move forward to analyze their interaction. Figure 5 shows these results with mean as an estimator.[12]

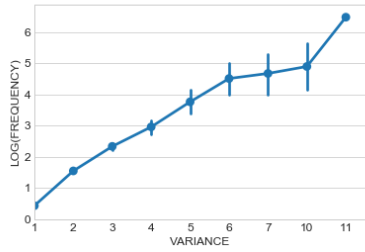
11 We expect that this skewness will also dominate system performance.

12 We observed comparable results with median as an estimator.

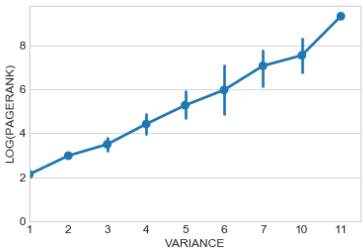
Figure 5: Correlations between head-tail phenomena (estimator=mean).



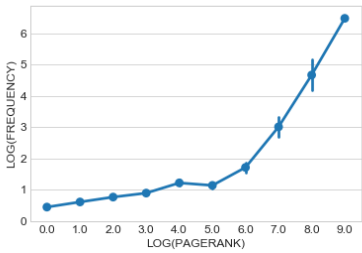
(a) Ambiguity and frequency.



(b) Variance and frequency.



(c) Variance and PageRank.



(d) PageRank and frequency.

Firstly, we predict a positive dependency between ambiguity of forms and their frequency (*D5*). We expect that frequently used forms tend to receive new meanings, often reflecting metonymy or meronymy of their dominant meaning. Figure 6a confirms this tendency, the Spearman correlation being 0.3772.

Secondly, we expect a positive correlation between variance of instances and their frequency (*D6*) or popularity (*D7*). Frequently mentioned and popular instances tend to be associated with more forms. Indeed, we observe that instances with higher frequency (Figure 6b) or PageRank (Figure 6c) typically have higher variance. The Spearman correlations are 0.6348 and 0.2542, respectively.

Thirdly, we compare the frequency of instances to their popularity measured with PageRank, predicting a positive correlation (*D8*). On average, this dependency holds (Figure 6d), though there are many frequent instances with low PageRank, or vice versa, leading to a Spearman correlation of 0.3281. The former are instances whose prominence coincides with the creation time of the corpus, but are not very well-known according to knowledge bases like DBpedia that were created at a later time point. As an example, consider the former Spanish politician and minister `Loyola_de_Palacio`, who was frequently in the news during the 90s, but today has a limited relevance and little available knowledge in Wikipedia/DBpedia. The latter are generally popular entities which were not captured sufficiently by the corpus, because their topical domain is marginal

to this corpus (e.g., scientists), or they became relevant after the corpus release (emerging entities).

All the correlations reported here (0.3772, 0.6348, 0.2542, and 0.3281) are positive. However, apart from the correlation between variance and frequency of instances, these correlations are not particularly high. The magnitude of these correlations can be explained with the fact that ambiguity and variance both have very coarse-grained values, and all aspects (ambiguity, variance, frequency, popularity) are heavily skewed towards a single value (of 1). As a result, there are many 'ties' during the computation of the correlations, which prohibits a clear trend to be visible through the Spearman correlations and makes these numbers difficult to interpret. We circumvent this by presenting plots where all Y-axis values for a single X-axis value are averaged (using mean or median). As discussed in the previous paragraphs, these plots show clear dependency between all four aspects we analyze here.

Hence, besides the high corpora skewness in terms of frequency, popularity, ambiguity, and variance, these factors also have positive interdependencies. Section 2.5 shows their effect on system performance.

2.4.6 *Frequency distribution for a single form or an instance*

We observed that the distribution of form frequency, instance frequency, and PageRank all have a Zipfian shape. But do we also see this behavior on a single form or instance level?

Supposedly, the frequency distribution within all forms that refer to an instance is Zipfian (*D9*). We test *D9* on the instance with highest variance and frequency, `United_States` (Figure 6). As expected, the vast majority of forms are used only seldom to refer to this instance, while in most cases a dominant short form "U.S." is used to make reference to this entity.

Figure 8a presents the frequency distribution of all instances that are referred to by the most ambiguous form in our data, "World cup". Figure 8b shows their PageRank distribution. In both cases, we observe a long-tail distribution among these instances (*D10*). Comparing them, we observe a clear match between frequency and PageRank, deviating only for instances that were prominent during the corpus creation, like `1998_FIFA_World_Cup`.

For analysis on the effect of frequency and popularity on system performance, please see Section 2.5.2.

Notably, we tested the hypothesis *D9* on a single instance (`United_States`), and the hypothesis *D10* on a single form ("World cup"). Whereas the obtained results fit our expectations, the sample size of 2 (1 form and 1 instance) is too small to make general conclusions about *D9* and *D10*. Unfortunately, the datasets which we analyze are still very limited in size, preventing us from testing these two hypotheses on a large scale.

Figure 6: Form frequencies for the instance `United_States`.

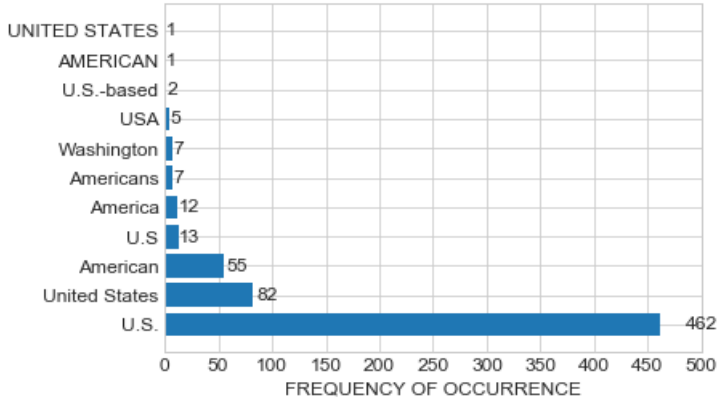
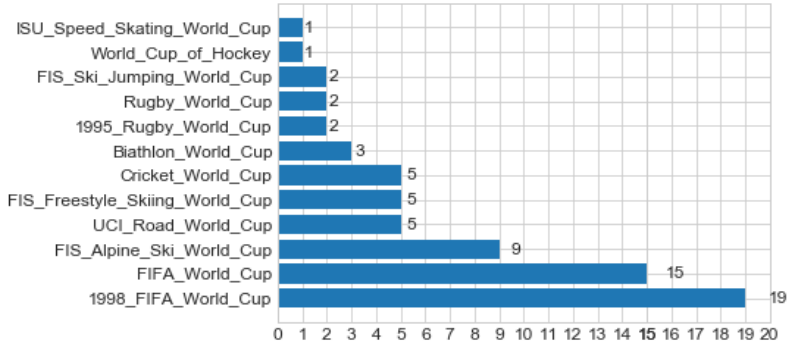
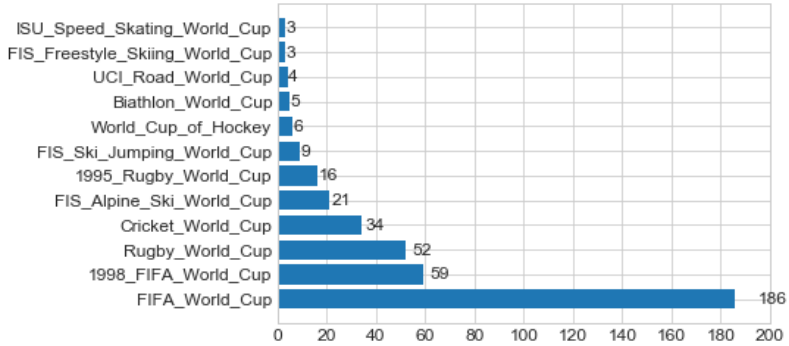


Figure 7: Distributions of the instances denoted by the most ambiguous form ("World Cup").

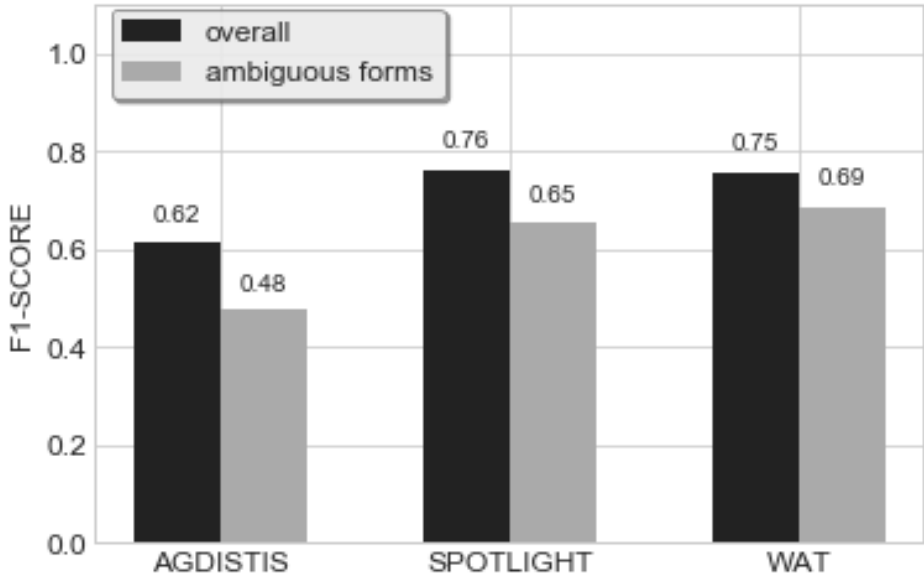


(a) Frequency distribution.



(b) PageRank distribution.

Figure 8: Micro F1-scores of systems: overall and on ambiguous subsets.



2.5 ANALYSIS OF SYSTEM PERFORMANCE AND DATA PROPERTIES

Next, we analyze system performance in relation to the data properties: ambiguity (Section 2.5.1), form frequency, instance frequency, and PageRank (Section 2.5.2), as well as their combinations (2.5.3 and 2.5.4).

2.5.1 *Correlating system performance with form ambiguity*

Figure 8 displays the micro F1-scores of AGDISTIS, Spotlight, and WAT on the two data collections jointly. For each system, we show its overall F1-score and F1-score on ambiguous forms only.

Section 2.4.3 showed that most of the forms in our corpora are not ambiguous. We expect that these forms lift the performance of systems, i.e., that they can resolve non-ambiguous forms easier than ambiguous ones (*S1*). Figure 8 confirms this for all systems: the F1-score on ambiguous forms is between 6 and 14 absolute points lower than the overall F1-score.

The F1-scores of the non-ambiguous forms for AGDISTIS, Spotlight, and WAT are 0.70, 0.77, and 0.78. Why is the score on the non-ambiguous forms not perfect? As explained in Section 2.3, in this chapter we consider the observed ambiguity as computed in corpora. Forms with an observed ambiguity of 1 can easily have much larger resource ambiguity, and even larger real-world ambi-

guity (real-world instances without a resource representation are manifested as NILs in EL).[13]

When computing macro- instead of micro-F1 scores, we observe similar findings for *S1*. Interestingly, the macro-F1 scores are consistently lower than the micro-F1 scores, especially in case of the ambiguous subsets evaluation. Namely, the overall macro-F1 scores are between 0.44 and 0.52, and between 0.14 and 0.34 on the ambiguous forms. This suggests that frequent forms boost system performance, especially on ambiguous surface forms. We investigate this further in the next Sections.

2.5.2 *Correlating system performance with form frequency, instance frequency, and PageRank*

	all forms			ambiguous forms only		
	FF-F1	**FI-F1**	**PR-F1**	**FF-F1**	**FI-F1**	**PR-F1**
AGDISTIS	0.2739	0.3812	0.1465	0.3550	0.4073	0.3969
Spotlight	0.1321	0.1847	0.1357	0.3986	0.4196	0.3108
WAT	0.4663	0.5050	0.3164	0.5831	0.5319	0.4214

Table 5: Correlation between F1-score and: frequency of forms (FF-F1), frequency of instances (FI-F1), and PageRank (PR-F1). Left: on all forms, right: only on ambiguous forms.

Next, we consider frequency of forms and instances, as well as PageRank values of instances in relation to system performance. For each of these, we expect a positive correlation with system performance (*S2*), suggesting that systems perform better on frequent and popular cases, compared to non-popular and infrequent ones.

The Spearman correlation for each of the systems and properties, over all forms, are shown in Table 5 (left half). While most of the correlation for frequency and popularity is positive, the values are in general relatively low (WAT being an exception). This shows that frequency/popularity by itself contributes, but is not sufficient to explain system performance. The right half of the Table shows the same metrics when applied to the ambiguous forms. We observe an increase in all values, which means that frequency and popularity are most relevant when multiple instances 'compete' sharing a form. These findings are in line with those in Section 2.5.1.

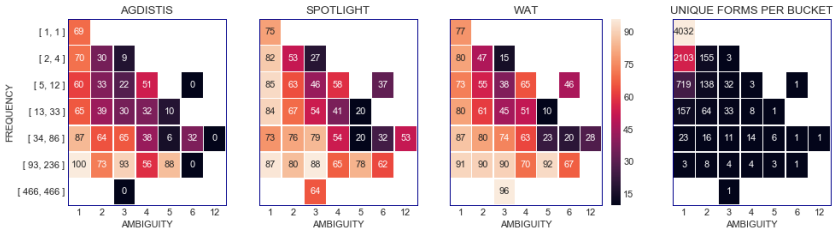


Figure 9: F1 per (ambiguity, frequency) pair. The first three heat maps show the F1-scores of systems per ambiguity degree and frequency range. Cells with higher F1-scores are assigned lighter colors. The last plot shows the amount of unique forms in each cell; here lighter coloring means higher frequency.

2.5.3 *Correlating system performance with ambiguity and frequency of forms jointly*

We have shown that system performance is lower on ambiguous than unambiguous forms. We also observed a tendency of systems to perform better on frequent forms and instances as compared to infrequent ones. But how does performance differ across different levels of ambiguity? How do ambiguity and form frequency interact as a joint predictor of performance?

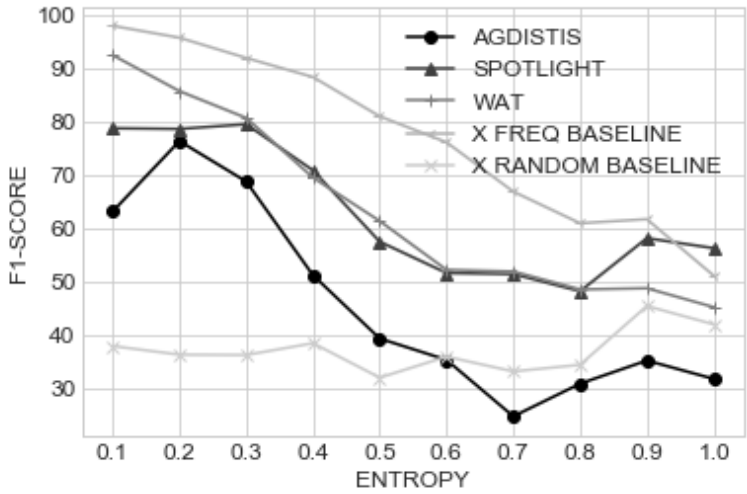
The heat maps in Figure 9 show the interplay between ambiguity, frequency, and micro F1-score for each of the systems. Generalizing over the discussion in Sections 2.5.1 and 2.5.2, we expect the best scores on frequent, less ambiguous forms (bottom-left), and worst F1-scores on infrequent, ambiguous forms (top-right) (*S3*). If we split the entire heat map along its inverse diagonal, we observe lighter colors (high accuracies) below this diagonal, and darker colors (low accuracies) above it. The high accuracies below the diagonal correspond to cases with high frequency and/or low ambiguity, whereas the low accuracies on the top correspond to cases with lower frequency and higher ambiguity, which confirms our hypothesis.

Note that the top-right corner of these heat maps is especially scarce, signaling the absence of ambiguous cases with a very low frequency. This can be explained with the hypothesis *D5* we posed in Section 2.3, on the positive correlation between frequency and ambiguity, or with the underlying Gricean maxims for informativeness and relevance of information. We expect a form to gain new meanings once the dominant instance it refers to becomes more frequent, whereas low-frequent forms are often non-ambiguous. The sparsity of the top-right corner is emphasized further with the small amount of forms with higher ambiguity in general in our corpora.

In Section 2.5.4, we investigate if some instances within ambiguous forms are more difficult than others.

13 We exclude the gold NIL entities from this analysis, but the systems still may predict a NIL entity for some of the non-NIL entities we consider.

Figure 10: Micro F1-score per entropy bucket.



2.5.4 *Correlating system performance with frequency of instances for ambiguous forms*

To measure the instance distribution within individual forms, we employ the notion of *normalized entropy*. The entropy of a form i with n_i instances and N_i occurrences is: $H_i = (-\sum_{j=1}^{n_i} p_{i,j} log p_{i,j})/log_2 N_i$, where $p_{i,j}$ is the probability that the instance j is denoted by the form i. For non-ambiguous forms $H_i = 0$, while forms with uniform frequency distribution of instances have a maximum $H_i = 1$. We predict an inverse correlation between system performance and entropy (*S4*). The results in Figure 10 show a dramatic drop in micro F1-score for uniformly distributed cases (high entropy) compared to skewed ones (low entropy). We compare these shapes to two baselines: frequency baseline, that picks the most frequent instance for a form on the gold data, and a random baseline, choosing one of the gold instances for a form at random. All three systems have a similar curve shape to the frequency baseline, whereas out of the three systems Spotlight's curve comes closest to that of the random baseline.

We also compute the macro F1-score per entropy bucket to help us understand whether the drop in performance in Figure 10 is due to: 1. a qualitative difference between low and high entropy forms, or 2. an overfitting of systems to the frequent interpretations of ambiguous forms. The macro F1-score reduces the effect of frequency on performance, by evaluating each form-instance pair once. We observe that the macro F1-scores are much more balanced across the entropy buckets compared to the micro F1-scores, and especially lower on the buckets with higher skewness (low entropy). This suggests that the high micro F1-score for low entropies is heavily based on frequent instances.

As a final analysis, we seek to understand whether frequent/popular instances of a form are indeed resolved easier than less frequent/popular instances of the same form. For that purpose, we pick the set of all ambiguous forms, and we rank their instances by relative frequency/PageRank value.

Considering the most ambiguous form "World Cup" as an example and ranking by frequency, its r1 (rank 1) instance is `1998_FIFA_World_Cup`, r2 is `FIFA_World_Cup`, ..., and r12 is `ISU_Speed_Skating_World_Cup`. We expect systems to perform much better on frequent instances of ambiguous forms, compared to infrequent instances of ambiguous forms, i.e., we expect F1-scores to decrease when moving from higher to lower ranks (*S5*). Figure 11 shows that our hypothesis holds to a large extent for precision, recall, and F1-scores, except for the occasional peaks at r6 and r9.

Figure 11: Precision, recall, and micro F1-score per instance frequency rank, averaged over the systems.

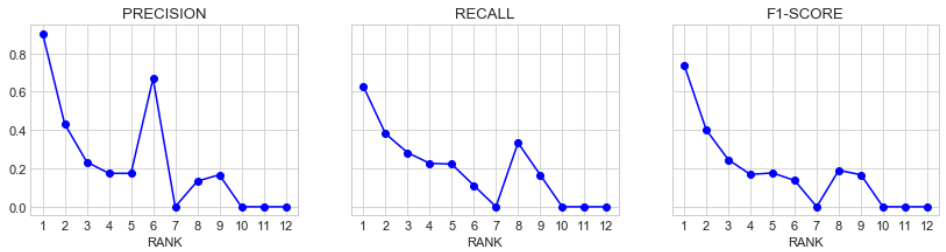
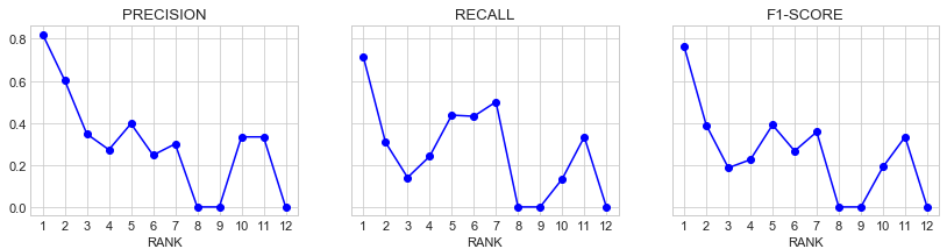


Figure 12: Precision, recall, and micro F1-score per PageRank-based rank, averaged over the systems.



Similarly, we order the instances denoting a form based on their relative PageRank value. We hypothesize that systems perform better on popular instances of ambiguous forms, compared to their unpopular instances (*S6*). Although less monotonic than the frequency ones in Figure 11, the resulting shapes of this analysis in Figure 12 suggest that popularity can also be applied to estimate system performance.

2.6 SUMMARY OF FINDINGS

We noted a positive correlation between ambiguity and frequency of forms, as well as between variance and frequency of instances. We noticed that the distribution of instances overall, but also per form, has a Zipfian shape. Similarly, the distribution of forms, both on individual and on aggregated level, is Zipfian. While some of these distributions are well-known in the community for words, this is the first time they have been systematically analyzed for surface forms of entities, their meaning and reference, and have empirically been connected with the performance of systems.

We observed that ambiguity of forms leads to a notable decline in system performance. Coupling it with frequency, we measured that low-frequent, ambiguous forms yield the lowest performance, while high-frequent, non-ambiguous forms yield the highest performance. Ambiguity here was approximated by the observed ambiguity, i.e., the ambiguity we measure in our corpora, which is expected to be far lower than the true ambiguity of a surface form.

The entropy of forms, capturing the frequency distribution of their denoted instances, revealed that balanced distributions tend to be harder for systems, with the micro F1-value dropping with 20-40 absolute points between the highest and lowest entropy. Finally, the higher performance on skewed cases was shown to be a result of overfitting to the most frequent/popular instances.

Based on these outcomes, we can conclude that the intersection of ambiguity and frequency/popularity is a good estimator of the complexity of the EL task. The difficult (tail) cases of EL should be sought among the low-frequent and unpopular candidates of ambiguous forms.

2.7 RECOMMENDED ACTIONS

We have shown that there are systematic differences between the head (E1) and the tail (E2) of the EL task, and that these reflect on how systems perform. Provided that systems show a weakness on tail cases, and that this weakness is simultaneously hidden by averaged evaluation numbers, how can we overcome this obstacle in practice? Here we list three recommendations:

1. When **creating a dataset**, we propose authors to include statistics on the head and the tail properties (ambiguity, variance, frequency, and popularity) of the data, together with a most-frequent-value baseline. By doing so, the community would be informed about the hard cases in that dataset, as well as about the portion of the dataset that can be resolved by following simple statistical strategies.

2. When **evaluating** a system, we suggest splitting of all cases into head and tail ones. Afterwards, head and tail cases can be evaluated separately, as well as together. This provides a direct insight into the differences in scoring of the tail cases compared to the head cases, potentially signaling aspects of the EL tail that are challenging for the given system. In addition,

the frequency skewness of head cases can be largely decreased by employing a macro instead of micro F1-score, as shown in this chapter.

3. In addition to the suggestion in 2., when **developing or training** a system, it should be made explicit which heuristics target which cases, and to what extent resources and training data optimize for the target dataset in relation to the head and tail distributions.

2.8 CONCLUSIONS

Although past research has argued that the performance of EL systems declines when moving from the head to the tail of the entity distribution, the long tail has not been quantified so far, preventing one to distinguish head (E1) and tail (E2) cases in the EL task. Previous linguistic studies on words distributions can also not be applied for this purpose since they do not study reference. We have hence set a goal to distinguish the head and the tail of the entity linking task (*RQ1*).

To achieve this, we performed the first systematic investigation into the relation of surface forms in EL corpora and instances in DBpedia. We provided a series of hypotheses that aim to explain the head and the tail of entity linking through data properties that capture this relation between surface forms and their instances. We analyzed existing EL datasets with respect to these properties, demonstrating that data properties have certain correlations that follow our hypotheses. Next, we investigated their effect on the performance of three state-of-the-art systems, proving that these data properties and their interaction consistently predict system performance. Namely, we noted a positive dependency of system performance on frequency and popularity of instances, and a negative one with ambiguity of forms. Our findings in this chapter are meant to influence future designs of both EL systems and evaluation datasets. To support this goal, we listed three recommended actions to be considered when creating a dataset, evaluating a system, or developing a system in the future.

We see two directions for future improvement of our analysis: 1. To obtain a corpus-independent inventory of forms and their candidate instances, both with their corresponding frequencies, is a challenge in the case of EL and no existing resource can be assumed to be satisfactory in this regard. We approximated these through the corpora we analyzed, but considering the fairly small size of most EL datasets, this poses a limitation to our current analysis. 2. Some of our current numbers are computed only on a handful of cases. This leads to unexpected disturbances in our results, like the occasional peaks for high ranks in Figure 11. We expect the outcome of this analysis to gain significance when more large EL datasets become available in the future.

Answer to RQ1 This chapter addresses the first research question of this thesis: *How can the long tail entities be distinguished from head entities?* We have shown that the head and the tail cases in the task of Entity Linking can be consistently distinguished along the data properties of ambiguity, frequency, and popularity. Furthermore, the observed dramatic decline of system performance from the head (low/no ambiguity, high frequency/popularity) towards the tail (higher ambi-

guity, low frequency/popularity) demonstrates the relevance of this head-tail distinction for subsequent research. Future designers of EL systems and datasets are advised to take the tail into account, and consult the recommended actions we listed in this chapter.

In the next part of the thesis, we analyze the representativeness of existing evaluation datasets for five semantic NLP tasks, including Entity Linking, for various aspects of the head and the tail. This aims at answering *RQ2*.

3 ANALYZING THE EVALUATION BIAS ON THE LONG TAIL OF DISAMBIGUATION & REFERENCE

Let us now turn to *RQ2: Are the current evaluation datasets and metrics representative for the long-tail cases?* Given the high overall accuracy scores reported by systems and the much lower accuracies on the tail cases (chapter 2), we expect that existing datasets are biased towards the head of the distribution and do not capture sufficiently the complex time-bound interaction between expressions and meanings discussed in chapter 1. In this chapter, we measure the representativeness of existing evaluation datasets through a set of metrics, and we propose approaches to improve it. We first present a preliminary study of the strengths and weaknesses of evaluation datasets in EL, and then expand it to provide a generic methodology applied to analyze five disambiguation and reference tasks, including EL. By doing so, we broaden our focus beyond the task of Entity Linking to also capture the tasks of Word Sense Disambiguation, Semantic Role Labeling, Event Coreference, and Entity Coreference.[1]

The content of this chapter is based on the research published in the following two publications:

1. Filip Ilievski, Marten Postma, and Piek Vossen (2016c). "Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text?" In: *The 26th International Conference on Computational Linguistics (COLING 2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1180–1191. URL: http://aclweb.org/anthology/C16-1112 Marten Postma and I collaboratively created the majority of the content of this paper. We worked together on all phases of the paper: motivation, literature review, methodology, implementation, and analysis, and only divided tasks within a phase.

2. Marieke Van Erp, Pablo N Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis (2016). "Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Vol. 5, p. 2016 I contributed to the formation of the idea of this paper, executed parts of its analysis, and wrote parts in various sections.

1 We use *meaning* as an umbrella term for both concepts and (event and entity) instances, and *lexical expression* as a common term for both lemmas and surface forms.

3.1 INTRODUCTION

Like many natural phenomena, the distribution of lexical expressions and their meanings follows a power law such as Zipf's law (Newman, 2005), with a few very frequent observations and a very long tail of low frequent observations.[2] This frequency distribution in our reporting about the world is guided by our interaction with the world and is not inherent to the physical existence of an entity or an event, since each (entity or event) instance either exists or it does not. In that sense, language and our writing about the world is heavily skewed, selective, and biased with respect to that world. We dominantly talk about only a few instances in the world and refer to them with a small set of expressions, which can only be explained by the contextual constraints within a language community, our social relationships, a topic, a location, and a period of time. Without taking these into account, it is impossible to fully determine meaning.

In chapter 1, we discussed the dependency of human language to these contextual constraints, especially with respect to time. Each text forms a unique semantic puzzle of expressions and meanings in which ambiguity is limited within the specific time-bound context, but is extreme without considering this context. The task of interpreting lexical expressions as meanings, known as disambiguation, has been addressed by the NLP community following a "divide & conquer" strategy that mostly ignores this complex time-bound relation.[3] Over the years, this resulted in numerous separate semantic tasks each with a specific set of datasets restricted to a small bandwidth with respect to the dynamics of the world and the large scope of the possible meanings that lexical expressions can have. By dividing the problem into different tasks on relatively small datasets, researchers can focus on specific subproblems and have their efforts evaluated in a straightforward manner. Datasets have been developed independently for each task, intended as a test bench to evaluate the accuracy and applicability of the proposed systems. Official evaluation scripts have been created for most datasets to enable a fair comparison across systems.

The downside of this practice is that task integration is discouraged, systems tend to be optimized on the few datasets available for each task, and the dependencies of ambiguities across tasks in relation to the time-bound contextual realities are not considered. As a result, there is little awareness of the overall complexity of the task, given language as a system of expressions and the possible interpretations given the changing world over longer periods of time. Systems are thus encouraged to strongly overfit on a single task, a single dataset, and a specific 'piece' of the world at a specific moment in time.

It is unclear whether the present manner of evaluating disambiguation and reference tasks is representative for the complex contextual dependency of the human language. Considering that the evaluation datasets have been created by sampling human communication which is dominated by a small set of very fre-

2 We acknowledge that there also exist many long-tail phenomena in syntactic processing, e.g., syntactic parsing.

3 While we believe that the other contextual constraints of topic, community, and location are also very relevant, in this chapter we focus on the temporal aspect of the disambiguation task.

quent observations, we hypothesize that the current evaluation exhibits a similar frequency bias. This bias towards the frequent and popular part of the world (*the head*) would be largely responsible for the performance of our disambiguation and reference tools. As evidenced in chapter 2, these tools are optimized to capture the head phenomena in text without considering the contextual constraints which are essential in order to deal with the less frequent phenomena in *the tail*. Therefore, their performance is expected to decline when moving from head towards tail cases. Understanding the evaluation bias better and improving its representativeness to challenge systems to perform well on the tail hence becomes a very important piece of the long-tail entities puzzle.

The main question we thus put forward and address in this chapter is: **are the current evaluation datasets and metrics representative for the long-tail cases?** We seek to understand to what extent do disambiguation and reference tasks cover the full complexity of the time-bound interaction between lexical expressions and meanings (in the broad sense of the word as defined here), and how can that be enhanced in the future.

We therefore first propose a number of metrics that formally quantity the complexity of this relation and apply this to a wide range of available datasets for a broad range of semantic tasks. Secondly, we provide evidence for the limitations of the current tasks and, thirdly, we present two proposals to improve these tasks in the hope that we challenge future research to address these limitations. We develop one of these proposals further in chapter 4, resulting in a semantic NLP task that is representative for the tail cases.

The chapter is structured as follows. We motivate the importance and relevance of this temporal interaction for both concept- and instance-based tasks in Section 3.2. Following up on previous research (Section 3.3) and our own preliminary study (Section 3.4), we define a model of the complex interaction (Section 3.5), and we conceptualize and formalize a collection of metrics in a generic manner (Section 3.6). Moreover, we apply these metrics to quantify aspects of existing evaluation sets (Section 3.7). In Section 3.8, we propose two approaches for creating metric-aware test sets that include a temporal dimension. The chapter is concluded in Section 3.9, where we also provide an answer to the *RQ2* from this thesis.

3.2 TEMPORAL ASPECT OF THE DISAMBIGUATION TASK

We live in a dynamic and rapidly changing world: some companies expand their offices all around the globe, while others collapse; people become celebrities overnight and are forgotten only several years afterwards. Similarly, a whole range of mainstream technological concepts of today's world have only been known since the last few decades. These observations have a big impact on the dynamics of a language system, since the relation between language expressions and meanings follows the changes in the world. To some extent this is reflected in new expressions and new meanings but most strongly this is reflected in the distributional usage of expressions and their dominant meaning.

Figure 13: Usage over time for two popular meanings that are referred to by the expression *Tesla*. The black line depicts the usage of *Tesla Motors Company*, while the gray line represents *Nikola Tesla*. The source of this data is Google Trends.

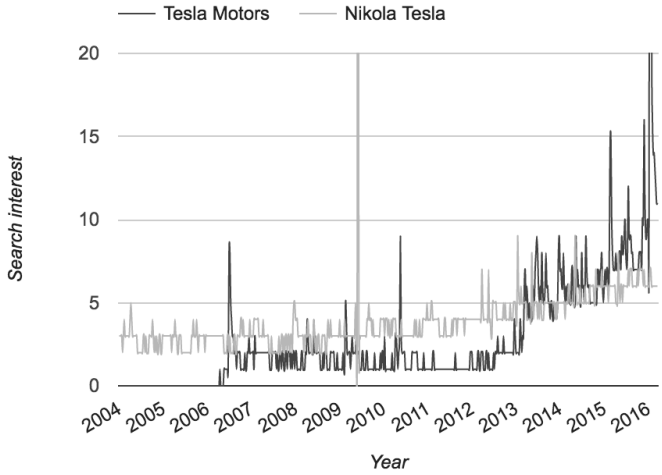
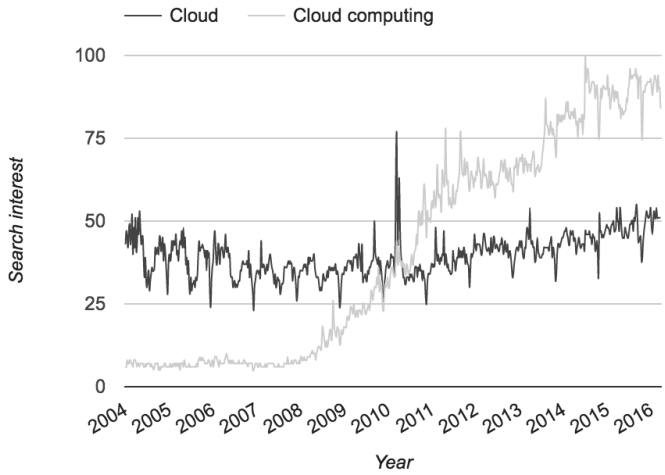


Figure 14: Usage distribution over time for two popular meanings referred to by the expression *cloud*. The black line depicts the usage of the clouds in the sky, natural objects placed in the atmosphere. The gray line stands for the modern meaning of cloud as an Internet-based type of computing. The source of this data is Google Trends.



For instance, the dominant meaning of the terms *mobile*, *cell*, and *phone* is the same for the contemporary, especially young, generations: mobile phone. On the other hand, older generations also remember different dominant concepts from the 80s and 90s: *mobile* being typically a decoration hanging from the ceiling, *cell* usually being a unit in a prison or body tissue, while *phone* referring to the

static devices found at home or on the streets. The dominant meanings of the 80s and 90s have been replaced by new dominant meanings, whereas the younger generation may have lost certain meanings such as the decoration. Similarly, football fans remember two different superstar *Ronaldo* players which have been dominant one after the other: the Brazilian striker and the Portuguese Ballon d'Or award winner.

What is shown by these examples is that not only new meanings appear and old meanings become obsolete but that, more strongly, the usage distribution of competing meanings changes over time. As the mobile phone gains popularity and the mobile decoration gets replaced by others, people refer to the mobile phone more often than the traditional mobile decoration. Hence, in a later point of time, the most commonly used meaning for *mobile* changes, even though both meanings are still possible. Similarly for the *Ronaldo* case: in 2016 one can still refer to both players, but the dominant meaning is now the Portuguese player.

We also observe a relation between the variety of lexical expressions used to refer to a meaning, and its dominance of usage. As the mobile phone gained popularity, its set of associated expressions expanded from only *mobile phone* to also: *mobile*, *phone*, *cell phone*, and *cell*. On the other hand, when referring to a prison cell, one should nowadays explicitly use the full expression *prison cell* instead of just *cell*, unless the surrounding context makes the reference of *cell* to *a prison cell* clear.

To measure the usage distribution of competing meanings, we could use online resources that track these distributions over time, such as Google Trends[4] and Wikipedia Views.[5] We present the usage distribution for instances denoted by *Tesla* in Figure 13, and for concepts expressed with the expression *cloud* in Figure 14. These plots demonstrate the ways in which the distribution of usage changes both for instances and concepts as a function of the temporal dimension.

As we discussed in section 3.1 and will be shown in our analysis in sections 3.4 and 3.7, the notion of time and its role in this mapping between expressions and meanings has not been taken into account in the creation of existing disambiguation and reference datasets. This observation points to a serious weakness in the representativeness of existing datasets for the full complexity of the disambiguation and reference tasks. The world we communicate about changes rapidly, leading to complex shifts in the meanings we refer to and the forms we use to make this reference. A static task of form-meaning mapping is hence too narrow compared to the rapidly changing world, discouraging systems to focus on the temporal aspect of the task. In reality, the same language system is still used for many different situations. While this works for humans, this is not yet solved for machines.

4 https://www.google.com/trends/

5 http://stats.grok.se/

3.3 RELATED WORK

The three problems enumerated in Section 3.1 have been addressed to some extent in past work.

Several approaches have attempted to resolve pairs of semantic NLP tasks jointly. Examples include: combined Entity Linking (EL) and Word Sense Disambiguation (WSD) (Hulpuş et al., 2015; Moro et al., 2014), combined event and entity coreference (EvC and EnC) (Lee et al., 2012) and resolving WSD and Semantic Role Labeling (SRL) together (Che and Liu, 2010). Although some task combinations are well-supported by multi-task datasets, such as CoNLL 2011 and 2012 for joint coreference (Pradhan et al., 2011, 2012), and (Moro and Navigli, 2015) for WSD and EL, still many multi-task systems have to be evaluated on separate datasets, each being a very small and independent sample of the communication about our world. Notable efforts to create multi-task annotated corpora are the AMR Bank (Banarescu et al., 2013) and the MEANTIME corpus (Minard et al., 2016).

Properties of existing datasets have been examined for individual tasks. For WSD, the correct sense of a lemma is shown to often coincide with the most frequent sense (Preiss, 2006) or the predominant sense (McCarthy et al., 2004). In the case of McCarthy et al. (2004), the predominant sense is deliberately adapted with respect to the topic of the text. Our work differs from (McCarthy et al., 2004) because they do not consider the temporal dimension. As a response to sense-skewed datasets, Vossen et al. (2013) created a balanced sense corpus in the DutchSemCor project in which each sense gets an equal number of examples. Cybulska and Vossen (2014) and Guha et al. (2015) both stress the low ambiguity in the current datasets for the tasks of EvC and EnC, respectively. Motivated by these findings, Guha et al. (2015) created a new dataset (QuizBowl), while Cybulska and Vossen (2014) extended the existing dataset ECB to ECB+, both efforts resulting in notably greater ambiguity and temporal diversity. As far as we are aware, no existing disambiguation/reference dataset has included the temporal dependency of ambiguity, variance, or dominance.

The Broad Twitter Corpus (Derczynski et al., 2016) represents a valuable attempt to capture temporal drift and spatial entity diversity in the genre of social media. To our knowledge, no similar effort exists for newswire.

In this chapter, we first investigate the representativeness of the EL datasets (section 3.4), measuring very little referential ambiguity. Evaluation is focused on well-known entities, i.e. entities with high PageRank (Page et al., 1999) values. Additionally, we observe a considerable overlap of entities across datasets, even for pairs of datasets that represent entirely different topics. Next, we systematically analyze datasets from the aforementioned five semantic NLP tasks in terms of their representativeness for the tasks they evaluate, and conclude similar semantic overfitting.

The problem of overfitting to a limited set of test data has been of central interest to the body of work focusing on domain adaptation (Carpuat et al., 2013; Daume III, 2007; Jiang and Zhai, 2007). In addition, unsupervised domain-

adversarial approaches attempt to build systems that generalize beyond the specifics of a given dataset, e.g., by favoring features that apply to both the train and target domains (Ganin et al., 2016). By evaluating on a different domain than the training one, these efforts have provided valuable insights into system performance. Nevertheless, this research has not addressed the aspects of time and location. Furthermore, to our knowledge, no approach has been proposed to generalize the problem of reference to unseen domains, which may be an artifact of the enormous amount of references that exist in the world leading to an almost infinite amount of possible classes to choose from.

We therefore propose to take this a step further and examine system performance with respect to a set of metrics, applicable over disambiguation tasks, thus setting the stage for creation of metric-aware datasets. We expect that these metrics show reduced complexity within well-defined temporal and topical boundaries and increased complexity across these boundaries. More extensive datasets than existing single- and multi-task datasets, driven by metrics on ambiguity, variance, dominance and time, would challenge semantic overfitting.

3.4 PRELIMINARY STUDY OF EL EVALUATION DATASETS

Before providing our methodology for analyzing disambiguation and reference datasets, we first present a preliminary study of the strengths and weaknesses of common benchmark datasets for EL. This section is based on the research published in (Van Erp et al., 2016).

EL benchmark datasets have often been treated as black boxes by published research, making it difficult to interpret efficacy improvements in terms of individual contributions of algorithms and/or labeled data. This scattered landscape of datasets and measures leads to a misleading interpretability of the experimental results, which makes the performance evaluation of novel approaches against the state-of-the-art rather difficult and open to several interpretations and questions. We thus inspected heterogeneous benchmark datasets in terms of genre such as newswire, blog posts, and microblog posts, through quantifiable aspects such as entity overlap, dominance, and popularity. This work is complementary to the overall analysis of the EL task by Ling et al. (2015b), to the research by Steinmetz et al. (2013) with focus on candidate generation, and to GERBIL (Usbeck et al., 2015), which aims to provide a central evaluation ecosystem for EL.

3.4.1 *Datasets*

Here we list, in an alphabetical order, the datasets that we analyzed.

AIDA-YAGO2 (Hoffart et al., 2011)[6] is an extension of the CoNLL 2003 entity recognition task dataset (Tjong Kim Sang and Meulder, 2003). It is based on news articles published between August 1996 and August 1997 by Reuters.

6 https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

Each entity is identified by its YAGO2 entity name, Wikipedia URL, and, if available, by Freebase Machine ID.

NEEL 2014 / 2015 The dataset of the #Microposts2014 Named Entity Extraction and Linking (NEEL) challenge (Cano et al., 2014)[7] consists of 3,504 tweets extracted from a much larger collection of over 18 million tweets. The tweets were provided by the Redites project, which covers event-annotated tweets collected for a period of 31 days between July 15th, 2011 and August 15th, 2011. It includes multiple noteworthy events, such as the death of Amy Winehouse, the London Riots, and the Oslo bombing. The 2015 corpus (Rizzo et al., 2015)[8] contains more tweets (6,025) and covers more noteworthy events from 2011 and 2013 (e.g., the Westgate Shopping Mall shootout), as well as tweets extracted from the Twitter firehose in 2014. The training set is built on top of the entire corpus of the #Microposts2014 NEEL challenge. It was further extended to include entity types and NIL references.

OKE2015 The Open Knowledge Extraction Challenge 2015 (Nuzzolese et al., 2015)[9] corpus consists of 197 sentences from Wikipedia articles. The annotation task focused on recognition (including coreference), classification according to the Dolce Ultra Lite classes,[10] and linking of named entities to DBpedia. The corpus was split into a train and a test set containing a similar number of sentences: 96 in the training set, and 101 in the test set.

RSS-500-NIF-NER (Röder et al., 2014)[11] contains data from 1,457 RSS feeds, including major international newspapers. The dataset covers many topics such as business, science, and world news. The initial 76-hour crawl resulted in a corpus which contained 11.7 million sentences. Out of these, 500 sentences were manually chosen to be included in the RSS500 corpus. The chosen sentences contain a frequent formal relation (e.g., "..who was born in.." for `dbo:birthPlace`).

WES2015 (Waitelonis et al., 2015) was originally created to benchmark information retrieval systems.[12] It contains 331 documents annotated with DBpedia entities. The documents originate from a blog about history of science, technology, and art.[13] The dataset also includes 35 annotated queries inspired by the blog's query logs, and relevance assessments between queries and documents. The WES2015 dataset is available as NIF2 dump[14], as well as in RDFa (Adida et al., 2012) format annotated within the HTML source of the blog articles.

7 http://scc-research.lancaster.ac.uk/workshops/microposts2014/challenge/index.html
8 http://scc-research.lancaster.ac.uk/workshops/microposts2015/challenge/index.html
9 https://github.com/anuzzolese/oke-challenge
10 http://stlab.istc.cnr.it/stlab/WikipediaOntology/
11 https://github.com/AKSW/n3-collection
12 http://yovisto.com/labs/wes2015/wes2015-dataset-nif.rdf
13 http://blog.yovisto.com/
14 http://s16a.org/node/14

WIKINEWS/MEANTIME [15] is a benchmark dataset that was compiled by the NewsReader project (Minard et al., 2016).[16] This corpus consists of 120 WikiNews articles, grouped in four sub-corpora with an equal size: Airbus, Apple, General Motors, and Stock Market. These are annotated with entities in text, including links to DBpedia, events, temporal expressions, and semantic roles. This subset of WikiNews news articles was specifically selected to represent entities and events from the domains of finance and business.

3.4.2 *Dataset characteristics*

Next we describe the characteristics of the analyzed benchmark datasets.

DOCUMENT TYPE The documents which comprise evaluation datasets can vary along several dimensions: 1. type of discourse/genre (news articles, tweets, transcriptions, blog articles, scientific articles, government/medical reports) 2. topical domain (science, sports, politics, music, catastrophic events, general/cross-domain) 3. document length (in terms of number of tokens: long, medium, short) 4. format (TSV, CoNLL, NIF; stand-off vs. in inline annotation) 5. character encoding (Unicode, ASCII, URL-encoding) 6. licensing (open, via agreement, closed).

Table 6 summarizes the document type characteristics of the corpora we analyze, already exposing notable diversity among the datasets with respect to the considered set of aspects.

Corpus	Type	Domain	Doc. Length	Format	Encoding	License
AIDA-YAGO2	news	general	medium	TSV	ASCII	Via agreement
NEEL 2014/2015	tweets	general	short	TSV	ASCII	Open
OKE2015	encyclopedia	general	long	NIF/RDF	UTF8	Open
RSS-500	news	general	medium	NIF/RDF	UTF8	Open
WES2015	blog	science	long	NIF/RDF	UTF8	Open
WikiNews	news	business	medium	stand-off XML	UTF8	Open

Table 6: General characteristics for analyzed datasets

15 We use the labels "WikiNews" and "MEANTIME" interchangeably to refer to this corpus for historical reasons.

16 http://www.newsreader-project.eu/results/data/wikinews

MENTION ANNOTATION CHARACTERIZATION When annotating entity mentions in a corpus, several either implicit or explicit decisions are being made by the dataset creators, that can influence evaluations on, and the comparison between, those datasets:

MENTION BOUNDARIES: inclusion of determiners ("the pope" vs "pope"), annotation of inner or outer entities ("New York's airport JFK" vs "JFK"), tokenization decisions ("New York's" vs "New York 's"), sentence-splitting heuristics.

HANDLING REDUNDANCY: annotating only the first vs. annotating all occurrences of an entity.

INTER-ANNOTATION AGREEMENT (IAA): one annotator vs multiple annotators, low agreement vs high agreement.

OFFSET CALCULATION: using 0 vs. using 1 as the initial identifier.

IRI VS. URI: character support for ASCII vs. Unicode.

NESTED ENTITIES: does the annotation allow for annotation of multiple entity layers, e.g., is 'The President of the United States of America' one entity in its entirety, or two disjoint entity mentions ('President' and 'United States of America'), or three mentions ('President', 'United States of America', 'The President of the United States of America')?

In the analyzed datasets, there is only limited variety on the entity boundaries and offsets, but each dataset was generated using different annotation guidelines, resulting in major differences between types of classes annotated, and which entities are (not) to be included. The 2014/2015 NEEL annotation guidelines, for example, are based on the CoNLL 2003 annotation guidelines (which also apply to AIDA-YAGO2) - however, while the CoNLL guidelines consider names of fictional characters as mentions of type `dbo:Person`, the NEEL guidelines consider this as a mention of type `dbo:FictionalCharacter`.

TARGET KNOWLEDGE BASE (KB) It is customary for the entity linking systems to link to cross-domain KBs: DBpedia, Freebase, or Wikipedia. Every dataset listed in Section 3.4.1 links to one of these general domain KBs. Almost all of these datasets refer to DBpedia, while AIDA-YAGO2 contains links to Wikipedia (which can easily be mapped to DBpedia) and Freebase. However, the cross-domain KBs lack coverage for entities that belong to the distributional tail. To evaluate entity linking on specific domains or non-popular entities, benchmark datasets that link to domain-specific resources of long-tail entities are required. We investigate the access and the potential for linking to such KBs in chapter 5.

	AIDA-YAGO2	NEEL2014	NEEL2015	OKE2015	RSS500	WES2015	Wikinews
AIDA-YAGO2 (5596)	-	327 (5.87)	451 (8.06)	0 (0)	70 (1.26)	269 (4.8)	65 (1.16)
NEEL2014 (2380)	327 (13.73)	-	1630 (68.49)	57 (2.39)	61 (2.56)	294 (12.35)	67 (2.82)
NEEL2015 (2800)	451 (16.11)	1630 (58.21)	-	56 (2)	71 (2.54)	222 (7.93)	72 (2.57)
OKE2015 (531)	0 (0)	57 (10.73)	56 (10.55)	-	13 (2.44)	149 (28.06)	21 (3.95)
RSS500 (849)	70 (8.24)	61 (7.18)	71 (8.36)	13 (1.53)	-	27 (3.18)	16 (1.88)
WES2015 (7309)	269 (3.68)	294 (4.02)	222 (3.04)	149 (2.04)	27 (0.16)	-	48 (0.66)
Wikinews (279)	65 (23.30)	67 (24.01)	72 (25.81)	21 (7.53)	16 (5.73)	48 (17.20)	-

Table 7: Entity overlap in the analyzed benchmark datasets. Behind the dataset name in each row the number of unique entities present in that dataset is given. For each datasets pair the overlap is given in number of entities and percentage (in parentheses).

3.4.3 *Distributions of instances and surface forms*

In this section, we analyze and compare the coverage of entities and entity mentions in the different datasets along three dimensions: entity overlap, entity distribution, and entity types.

ENTITY OVERLAP In Table 7, we present the entity overlap between the different benchmark datasets. Each row in the table represents the percentage of unique entities present in that dataset that are also represented in the other datasets. As the table illustrates, there is a fair overlap between the entities in the WikiNews dataset and the other benchmark datasets. The overlap between the NEEL2014 and NEEL2015 datasets is explained by the fact that the latter is an extension of the former. The WES2015 dataset has the least in common with the other datasets.

AMBIGUITY Let the true ambiguity of a surface form s be the number of meanings that this surface form can have. As new places, organizations and people are named every day, without access to an exhaustive collection of all named entities in the world, the true ambiguity of a surface form is unknown. However, we can estimate the ambiguity of a surface form through the function $A(s) : S \rightarrow \mathbb{N}$ that maps a surface form to an estimate of the size of its candidate mapping, such that $A(s) = |C(s)|$. Estimating the ambiguity by a lookup in a resource, like DBpedia, equals the resource ambiguity which we defined in chapter 2.

The ambiguity of, for example, a place name offers only a rough *a priori* estimate of how difficult it may be to disambiguate that surface form. Observation of annotated occurrences of this surface form in a text collection (observed ambiguity, chapter 2) allows us to make more informed estimates. In Table 8, we show the average number of meanings denoted by a surface form, indicating the ambiguity, as well as complementary statistical measures on the datasets. In this Table, we observe that most datasets have a low number of average meanings per surface form. The standard deviation varies across datasets. In particular, the OKE2015 and WikiNews/MEANTIME datasets stand out in their high number of maximum meanings per surface form and standard deviations.

Given a surface form, some senses are much more *dominant* than others – e.g., for the name 'Berlin', the resource dbpedia:Berlin (Germany) is much more 'talked about' than `Berlin,_New_Hampshire` (USA). Therefore, we also take into account estimates of *prominence / popularity* and *dominance*.

POPULARITY Let the true popularity of a resource r_i be the percentage of other resources $r_k \in R$, which are less known than r_i. Let the popularity estimate $Pr(r_i)$ be the relative frequency with which the resource r_i appears linked on Wikipedia compared to the frequency of all other resources in R. Formally:

$$Pr(r_i) = \frac{\sum_{s \in S} |WikiLinks(s, r_i)|}{\sum_{s \in S, r \in R} |WikiLinks(s, r)|}$$

Corpus	Average	Min.	Max.	σ
AIDA-YAGO2	1.08	1	13	0.37
2014 NEEL	1.02	1	3	0.16
2015 NEEL	1.05	1	4	0.25
OKE2015	1.11	1	25	1.22
RSS-500	1.02	1	3	0.16
WES2015	1.06	1	6	0.30
WikiNews	1.09	1	29	1.03

Table 8: Ambiguity statistics for analyzed datasets. Average stands for average number of meanings per surface form, Min. and Max. stand for the minimum and maximum number of meanings per surface form found in the corpus respectively, and σ denotes the standard deviation. We note that the number of meanings includes NILs.

As in chapter 2, we estimate entity popularity through PageRank (Page et al., 1999). Some entities which are linked from only a few, but very prominent entities are also considered popular. *Goethe's Faust*, for example, only has a few links, but one of those is *Goethe*, which is considered a prominent entity, and thus, *Goethe's Faust* would also be popular.

Figure 15 depicts the PageRank distribution of the DBpedia based benchmarks compared to each other, as well as compared to the overall PageRank distribution in DBpedia.[17] The figure illustrates that all investigated benchmarks favor entities that are much more popular than average entities in DBpedia. Thus, the benchmarks show a considerable bias towards head entities. However, the whiskers of the box plots also show that all benchmarks contain long-tail entities (i.e., all benchmarks contain some entities with minimum PageRank), and almost all of them also contain the DBpedia entity with the highest PageRank value (i.e., `United_States`).

Evaluating against a corpus with a tendency to focus strongly on prominent (popular) entities may cause some issues regarding the applicability of the developed systems. Entity Linking systems that include the global popularity of entities in their approach can reach very good results (Tristram et al., 2015), but these can hardly be transferred to other settings in which these popular entities do not prevail.

DOMINANCE Let the true dominance of a resource r_i for a given surface form s_i be a measure of how commonly r_i is meant with regard to other possible meanings when s_i is used in a sentence. Let the dominance estimate $D(r_i, s_i)$

17 We use the DBpedia PageRank from `http://people.aifb.kit.edu/ath/`.

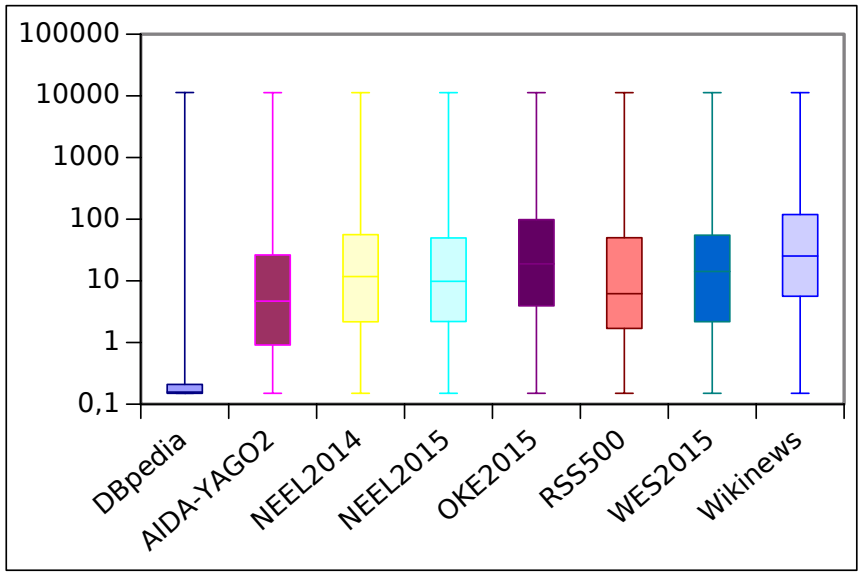


Figure 15: Distribution of DBpedia entity PageRank in the analyzed benchmarks. The leftmost bar shows the overall PageRank distribution in DBpedia as a comparison. The boxes depict the PageRank of the 25% of the instances with a PageRank above and below the median, respectively, while the whiskers capture the full distribution.

be the relative frequency with which the resource r_i appears in Wikipedia links where s_i appears as the anchor text. Formally:

$$D(r_i, s_i) = \frac{|WikiLinks(s_i, r_i)|}{\forall_{r \in R} |WikiLinks(s_i, r)|}$$

The dominance statistics for the analyzed datasets are presented in Table 9. The dominance scores for all corpora are quite high and the standard deviation is low, meaning that in vast majority of cases, a single instance is associated with a certain surface form in the annotations. More statistics for dominance can be found on the GitHub page of this study.[18]

Corpora that contain resources with high ambiguity, low dominance, and low popularity can be considered more difficult to disambiguate. This is due to the fact that such corpora require a more careful examination of the context of each mention before algorithms can choose the most likely disambiguation. In cases with low ambiguity, high prominence, and high dominance, simple popularity-based baselines that ignore the context of the mention can already perform quite accurately.

ENTITY TYPES Entities characterized with certain semantic types may be more difficult to disambiguate than others. For example, while country and company

18 The entire code of the analysis reported in this study can be found on GitHub: `https://github.com/dbpedia-spotlight/evaluation-datasets`.

Corpus	Dominance	Max	Min	σ
AIDA-YAGO2	0.98	452	1	0.08
NEEL 2014	0.99	47	1	0.06
NEEL 2015	0.98	88	1	0.09
OKE2015	0.98	1	1	0.11
RSS-500	0.99	1	1	0.07
WES2015	0.97	1	1	0.12
WikiNews	0.99	72	1	0.09

Table 9: Dominance statistics for analyzed datasets.

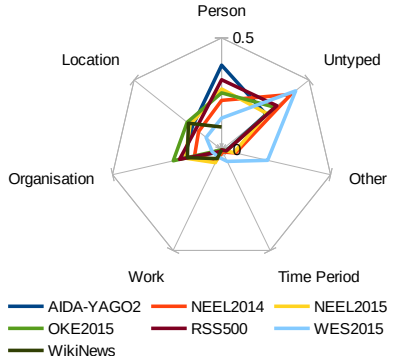
names (e.g., *Japan*, *Microsoft*) are more or less unique, names of cities (e.g., *Springfield*) and persons (e.g., *John Smith*) are generally more ambiguous. Thus, we can expect that the distribution of entity types has a direct impact on the difficulty of the entity linking task.

We analyzed the types of entities in DBpedia with respect to our benchmark datasets. For that analysis, we used RapidMiner[19] with the Linked Open Data extension (Ristoski et al., 2015). Figure 16 shows the overall distribution, as well as a breakdown by the most frequent top classes. Although types in DBpedia are known to be notoriously incomplete (Paulheim and Bizer, 2014), and NIL entities are not considered, these figures still reveal some interesting characteristics:

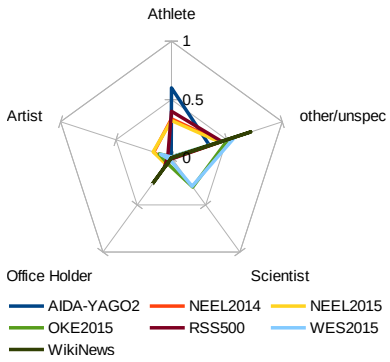
- AIDA-YAGO2 has a tendency towards sports related topics, as shown in the large fraction of sports teams and athletes.
- NEEL2014 and WES2015 treat time periods (i.e., years) as entities, while the others do not.
- OKE2015 and WES2015 have a tendency towards science-related topics, as shown in the large fraction of the entity types `Scientist` and `EducationalInstitution` (most of the latter are universities).
- WikiNews/MEANTIME, without surprise, has a strong focus on politics and economics, with a large portion of the entities being of classes `OfficeHolder` (i.e., politicians) and `Company`.
- The WES2015 corpus has a remarkably larger set of *other* and *untyped* entities. While many corpora focus on persons, places, etc., WES2015 also expects annotations for general concepts, e.g., `Agriculture` or `Rain`.

These findings reveal that it is difficult to build NER/EL tools that perform well on all these datasets. For instance, for most datasets annotations of general concepts would be punished as false positives, whereas WES2015 would expect them and punish their absence as false negatives.

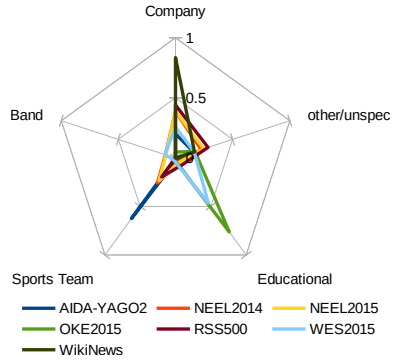
19 http://www.rapidminer.com/



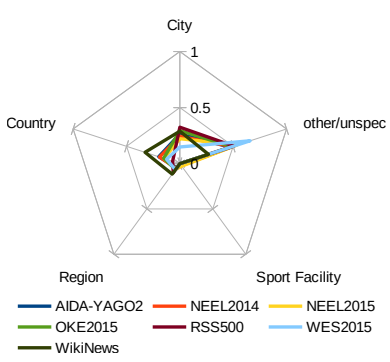
(a) Overall distribution



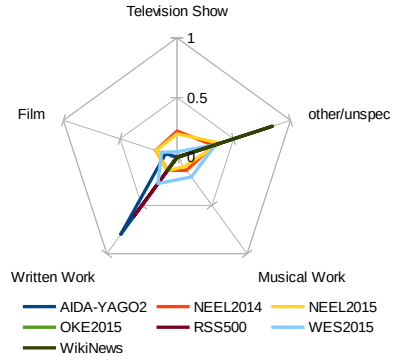
(b) Distribution of person subtypes



(c) Distribution of organization subtypes



(d) Distribution of location subtypes



(e) Distribution of work subtypes

Figure 16: Distribution of entity types overall (a), as well as a breakdown for the four most common top classes person (b), organization (c), location (d), and work (e). The overall distribution depicts the percentage of DBpedia entities (a), the breakdowns depict the percentages in the respective classes.

3.4.4 *Discussion and roadmap*

A number of benchmark datasets for evaluating entity linking exist, but our analyses show that these datasets suffer from two main problems:

INTEROPERABILITY AND EVALUATION Dataset evaluation and interoperability between datasets are far from trivial in practice. Existing datasets were created by following different annotation guidelines (mention and sentence boundaries, inter-annotation agreement, etc.) and have made different choices regarding their implementation decisions, including encoding, format, and number of annotators. These differences make the interoperability of the datasets and the unified comparison between entity linking systems over these datasets difficult, time-consuming, and error-prone.

POPULARITY AND NEUTRAL DOMAIN Although the datasets claim to cover a wide range of topical domains, in practice they seem to share two main drawbacks: skewness towards popularity and frequency, and coverage of well-known entities from a neutral domain.

While it is not possible to find a 'one-size-fits-all' approach to creating benchmark datasets, we do believe it is possible to define directions for general improvement of benchmark datasets for entity linking.

DOCUMENTATION Seemingly trivial choices such as initial offset count, or inclusion of whitespace can make a tremendous difference for the users of a dataset. This is applicable to any decision made in the process of creation of a dataset. Until these considerations become standardized in our community, we advise dataset creators to document their decisions explicitly and in sufficient detail.

STANDARD FORMATS As annotation formats are becoming more standardized, dataset creators have more incentive to choose an accepted data format, or provide script that converts the original data to one or more standardized formats.

DIVERSITY The majority of the datasets we analyzed link to generic KBs and focus on prominent entities. To gain insights into the usefulness of entity linking approaches and understand their behavior better, we need to evaluate these on datasets characterized with high ambiguity, low dominance, and low popularity. For this purpose, we also need datasets that focus on long-tail entities and different domains. We note that such datasets would be created by sampling existing communication in an 'unnatural' way, by deliberately picking text documents that describe long-tail instances and that exhibit certain desired properties (like low dominance or high ambiguity). By doing so, we would challenge systems to also perform well on tail cases and seek to inspire methods that avoid overfitting to the head of the distribution. We present an example for such a task and dataset in chapter 4 of this thesis.

To summarize, this study reveals two main problems with current evaluation in EL: interoperability of representation formats and representativeness of content. The former challenge could potentially be addressed with methods from

the Linked Data community; however, we will not dive further into this discussion as it is outside of the scope of this thesis.[20]

We focus on the challenge of dataset representativeness in the remainder of this chapter. Our approach is based on a diverse set of generic metrics that can be used to measure the representativeness of any dataset that evaluates disambiguation or reference in NLP.

Before describing the metrics, in the next section we first define a model that captures the interaction between the lexical expressions found in text and the meanings in the world.

3.5 SEMIOTIC GENERATION AND CONTEXT MODEL

We want to model the relation between expressions and meanings in the world within a generation that shares the same language system, as well as the fluctuation in usage of expressions and meanings over time within this generation. We therefore assume that for each language community at a specific time, there exist a set of meanings M in the world and a set of lexical expressions L in a language. The relation between these sets is many-to-many: each lexical expression L_i can refer to multiple meanings $M_1, M_2, ...$ (ambiguity) and each meaning M_j can be verbalized through multiple lexical expressions $L_1, L_2, ...$ (variance). As we discuss in Section 3.2, the sets of M, L, their relations, and especially the distributions of these relations, are dynamic, i.e. they can change over time. We call this model "the Semiotic generation and context model", because it captures the distribution changes in the semiotic relation between meanings and lexical expressions, given the context of the changes in the world and within the language system of a generation.

In practice, we study available proxies of the world at a moment in time and of the language of a generation which capture this relation at a given time snapshot: lexical resources are considered as a proxy of the language system of a generation and the dataset is considered as a proxy for the world at a particular moment in time creating a specific context. We analyze the time-anchored interaction between M and L in the datasets proxy and measure this against their interaction in the resources proxy to provide insight on how representative the datasets are for the task. Note that the proxies of datasets and resources cover only a subset of the language used within a generation, and (consequently) only a subset of all possible meanings. While not ideal, this is the best we have because there is no way to capture all language used within a generation nor list every possible meaning, especially considering that we can always create new meanings, e.g., by inventing some non-real world ones.

20 We refer the reader to GERBIL's evaluation framework for a thorough attempt to reconcile the different formats used by EL datasets and systems.

3.6 METHODOLOGY

Based on the Semiotic generation and context model, we now define and formalize a number of metrics that qualify datasets for disambiguation and reference tasks. In this Section, we describe these metrics and explain the tasks we focus on. Furthermore, we enumerate the design choices that guide our pick of datasets and we elaborate on the datasets we analyze.

3.6.1 *Metrics*

MEAN OBSERVED AMBIGUITY (MOA) As in chapter 2, we define observed ambiguity of an expression as the cardinality of the set of meanings it refers to within a dataset (O_{L_i}). For example, the expression *horse* has 4 meanings in WordNet but only the chess meaning occurs in the dataset, resulting in an observed ambiguity of 1. The Mean Observed Ambiguity (MOA) of a dataset is then the average of the individual observed ambiguity values.[21]

MEAN OBSERVED VARIANCE (MOV) We define observed variance of a meaning as the cardinality of the set of lexical expressions that express it within a dataset (O_{M_j}). The chess meaning of *horse* also has *knight* as a synonym but only *horse* occurs in the dataset, hence an observed variation of 1. The Mean Observed Variance (MOV) of a dataset is then the average of the individual observed variance values.

MEAN OBSERVED DOMINANCE OF AMBIGUITY (MODA) We define dominance of ambiguity as a frequency distribution of the dominant meaning of a lexical expression. For example, *horse* occurs 100 times in the data and in 80 cases it has the chess meaning: the dominance score is 0.8. The Mean Observed Dominance of Ambiguity (MODA) of a dataset is the average dominance of all observed expressions.[22]

MEAN OBSERVED DOMINANCE OF VARIANCE (MODV) We define the notion of dominance of variance, as a frequency distribution of the dominant lexical expression referring to a meaning. If *horse* is used 60 times and *knight* 40 times for the same meaning then the observed dominance of variance is 0.6. The Mean Observed Dominance of Variance (MODV) of a dataset is then the average dominance computed over all observed meanings.

ENTROPY OF THE MEANINGS (NORMALIZED) OF A L.E. (EMNLE) We define an alternative notion of dominance, based on entropy, in order to consider the distribution of the less dominant classes in a dataset. We introduce $p(M_j|L_i)$: a conditional probability of a meaning M_j based on the occurrence of a lexical expression L_i. We compute this probability using the formula $p(M_j|L_i) = \frac{p(M_j,L_i)}{p(L_i)}$, a ratio between the number of common

21 This metric corresponds to the 'average' (ambiguity) column in Table 3.

22 This metric corresponds to the 'dominance' column in Table 9.

occurrences of M_j and L_i, and on the other hand, the total number of occurrences of L_i regardless of its meaning. We combine the individual conditional probabilities for L_i in a single information theory metric of entropy, $H(O_{L_i})$:

$$H(O_{L_i}) = \frac{-\sum_{j=1}^{n} p(M_j|L_i)\log_2 p(M_j|L_i)}{\log_2(n)} \tag{1}$$

For example, given 100 occurrences of the lexical expression *horse*, where 80 occurrences refer to the the chess meaning and 20 to the animal meaning, the entropy of the expression *horse* would be 0.72. To compute a single entropy (EMNLE) value over all lexical expressions in a dataset, we average over the individual entropy values:

$$EMNLE(O_L, R_L) = \frac{1}{n}\sum_{i=1}^{n} H(O_{L_i}, R_{L_i}) \tag{2}$$

ENTROPY OF THE L.E.S (NORMALIZED) OF A MEANING (ELENM)

We introduce $p(L_i|M_j)$: a conditional probability of a lexical expression L_i based on the occurrence of a meaning M_j. We compute this probability using the formula $p(L_i|M_j) = \frac{p(L_i,M_j)}{p(M_j)}$, a ratio between the number of common occurrences of M_j and L_i, and on the other hand, the total number of occurrences of M_j regardless of its expression. We combine the individual conditional probabilities for M_j in a single information theory metric of entropy, $H(O_{M_j})$:

$$H(O_{M_j}) = \frac{-\sum_{i=1}^{n} p(L_i|M_j)\log_2 p(L_i|M_j)}{\log_2(n)} \tag{3}$$

Suppose the meaning of horse as a chess piece is expressed 60 times by the lexical expression *horse* and 40 times by *knight*, then the entropy of the chess piece meaning of *horse* is 0.97. To compute a single entropy (ELENM) value over all meanings in a dataset, we average over the individual entropy values:

$$ELENM(O_M, R_M) = \frac{1}{n}\sum_{j=1}^{n} H(O_{M_j}, R_{M_j}) \tag{4}$$

RELATION BETWEEN OBSERVED & RESOURCE AMBIGUITY (RORA) We define resource ambiguity of a lexical expression as the cardinality of the set of meanings that it can refer to according to a lexical resource (R_{L_i}). Then we define the ratio between observed and resource ambiguity for a lexical expression as:

$$ratio_{amb}(O_{L_i}, R_{L_i}) = \frac{|\{M_j : M_j \in O_{L_i}\}|}{|\{M_j : M_j \in R_{L_i}\}|} \tag{5}$$

In the case that only 1 out of 4 resource meanings is observed in the dataset, for example only the chess meaning of *horse*, this would lead to a $ratio_{amb}$ value of 0.25. To compute the RORA value of a dataset, we average over the individual ratios:

$$RORA(O_L, R_L) = \frac{1}{n} \sum_{i=1}^{n} ratio_{amb}(O_{L_i}, R_{L_i}) \tag{6}$$

RELATION BETWEEN OBSERVED AND RESOURCE VARIANCE (RORV) We define resource variance of a meaning as the cardinality of the set of lexical expressions which can verbalize it (R_{M_j}). Then we define the ratio between observed and resource variance for a given meaning:

$$ratio_{var}(O_{M_j}, R_{M_j}) = \frac{|\{L_i : L_i \in O_{M_j}\}|}{|\{L_i : L_i \in R_{M_j}\}|} \tag{7}$$

Suppose that the expressions *horse* and *knight* can refer to the meaning of chess piece according to a resource, but only the expression *horse* refers to it in a particular dataset, this would lead to a $ratio_{var}$ value of 0.5. To compute the RORV value of a dataset, we average over the individual ratios:

$$RORV(O_M, R_M) = \frac{1}{n} \sum_{i=1}^{n} ratio_{var}(O_{M_j}, R_{M_j}) \tag{8}$$

AVG TIME-ANCHORED RANK (ATR) Since the relevance of meanings is not constant over time, we define the popularity of a meaning in a point of time, $popularity_{M_j}(t)$. A lexical expression can potentially denote multiple meanings, each characterized with a certain degree of time-anchored popularity. Likewise, we order the list of candidate meanings for a given lexical expression based on their popularity at the moment of publishing of the dataset document. For example, if the dataset covers news about a chess tournament, we will see a temporal peak for the chess meaning of *horse* relative to the other meanings. The popularity rank of each meaning, including the correct gold standard meaning, is its position in this ordered list. By averaging over the ranks of all golden candidates we can compute the Average Time-anchored Rank of the golden candidates in a dataset, which gives an indication about the relation between the relative temporal popularity of a meaning and the probability that it is the correct interpretation of an expression, varying from stable to extremely dynamic relations. An ATR rank of a dataset close to 1 indicates a strong bias towards the popular meanings at the time of creation of the dataset.

AVG TIME-ANCHORED RELATIVE FREQUENCY OF USAGE (ATRFU) The potential bias of meaning dominance with respect to its temporal popularity can alternatively be assessed through its frequency of usage at a

point of time. We denote the usage of a meaning with U_{M_j}. For a given lexical expression, we compute the relative temporal frequency of usage (FU) of the golden meaning relative to the frequency of usage of all candidate meanings:

$$FU_{M_j}(t) = \frac{U_{M_j}(t)}{\sum_{i=1}^{n} U_{M_i}(t)} \tag{9}$$

The average relative frequency of usage at a given time point (ATRFU) is an average of the frequency values of all gold standard meanings in a dataset. We introduce this metric in order to gain insights into the popularity difference between the competitive meanings at a given time period. This metric would allow us, for instance, to detect that in July 2014 *the United States men's national soccer team* was much more popular than *the women's national soccer team*, while *Tesla Motors* was only slightly more popular than *Nikola Tesla* in May 2015.

DATASET TIME RANGE (DTR) We define DTR as a time interval between the earliest and the latest published document of a dataset:

$$DTR = [min(date_{doc}), max(date_{doc})] \tag{10}$$

where $date_{doc}$ is the publishing date of a document. For instance, the DTR of the MEANTIME (Minard et al., 2016) dataset is $[2004, 2011]$.

3.6.2 *Tasks*

We demonstrate the applicability of the metrics defined in Section 3.6.1 on a selection of disambiguation and reference tasks. We cover both instance-based tasks (EL, EnC, and EvC), as well as concept-oriented tasks (WSD and SRL).[23] In Table 10, we specify the model components per task, enabling the metrics to be computed. The metrics concerning lexical resources (WordNet (Fellbaum, 1998) for WSD, and PropBank (Kingsbury and Palmer, 2002) for SRL) are only computed for the concept-oriented tasks. Whereas lexical resources, such as WordNet and PropBank, can be seen as reasonable proxies for most of the expressions and concepts known to a generation, it is more difficult to consider databases of instances, such as DBpedia,[24] to approximate all the possible instances that expressions, e.g., *Ronaldo*, can refer to. This is especially the case for events, e.g. the goals *Ronaldo* scored, or the *Ronaldo* t-shirts being sold in a fan shop. There is hardly any registry of real world events independent of the mentions of events in text. Likewise, we only find a few *Ronaldo* entities in DBpedia. Despite its impressive size, DBpedia only covers a very small subset of all instances in the world.

23 Note that in the case of SRL we focus on the expression-to-meaning mapping of predicates and do not analyze roles.

24 `http://dbpedia.org`

Task	Lexical expression	Meaning	Resource
EL	entity mention	entity	DBpedia
EnC	entity mention	entity	DBpedia
EvC	event mention	event	/
WSD	lemma	sense	WordNet
SRL	predicate mention	predicate	PropBank

Table 10: Task specification of model components.

3.6.3 *Datasets*

The choice of datasets conforms to the following rationale. We consider test datasets with running text in English,[25] because we assume that they are the most natural instantiations of the interaction between lexical expressions and meanings and tend to report on the changes in the world. Moreover, such datasets lend themselves better for joint tasks. Finally, we favor publicly available datasets which are commonly used in recent research. The chosen datasets per task are as follows.

EL We consider the following datasets: AIDA-YAGO2 (**AIDA test B**) (Hoffart et al., 2011), **WES2015** (Waitelonis et al., 2015), and **MEANTIME** (Minard et al., 2016). We analyze the commonly used test B collection from the AIDA-YAGO2 dataset, which contains 5,616 entity expressions in 231 documents. WES2015 contains 13,651 expressions in 331 documents about science, while the MEANTIME corpus consists of 120 documents regarding four topics, with 2,750 entity mentions in total.

ENC Guha et al. (2015) created a dataset, QuizBowl, for nominal coreference, containing 9,471 mentions in 400 documents. The data annotated comes from a game called quiz bowl.[26]

EVC we consider three event coreference corpora: EventCorefBank (**ECB**) (Lee et al., 2012), **ECB+** (Cybulska and Vossen, 2014), and EventNuggets (**TAC KBP '15**) (Mitamura et al., 2015b). ECB contains 480 documents spread over 43 topics, while its extension ECB+ contains an additional 502 documents spread over the same set of topics. The training corpus of TAC KBP '15 contains 7,478 event coreference chains (hoppers).[27]

25 Our analysis in this chapter is performed on 13 English datasets. The metrics we define in Section 3.6.1 can easily be applied to many other languages. Namely, the resource-dependent metrics (RORA and RORV) can be applied to the wide range of languages in which DBpedia/WordNet/PropBank are available (for an illustration, DBpedia is currently available in 125 languages). Furthermore, all other metrics rely solely on the annotated textual content within a corpus, which makes them applicable for any language.

26 `https://en.wikipedia.org/wiki/Quiz_bowl`

27 We were unable to obtain the test data for the TAC KBP '15 dataset, hence our analysis is performed on the training data.

WSD The following datasets were taken into consideration: Senseval–2 (**SE2 AW**): All-Words task (Palmer et al., 2001) ; Senseval-3 (**SE3 task 1**): Task 1: The English all-words task (Snyder and Palmer, 2004) ; SemEval-2007 (**SE7 task 17**): Task-17: English Lexical Sample, SRL and All Words (Pradhan et al., 2007) ; SemEval–2010 (**SE10 task 17**): Task 17: All-Words Word Sense Disambiguation on a Specific Domain (Agirre et al., 2010); SemEval–2013 (**SE13 task 12**): Task 12: Multilingual Word Sense Disambiguation (Navigli et al., 2013). The number of test items per competition ranges from roughly 500 to 2,500 instances.

SRL For Semantic Role Labeling, we selected the CoNLL-2004 Shared Task: Semantic Role Labeling (**CoNLL04**) (Carreras and Màrquez, 2004). In total, 9,598 arguments were annotated for 855 different verbs.

3.7 ANALYSIS

In this Section, we measure to what extent datasets cover the complexity of the task they evaluate.[28]
According to Table 11, high complexity in both directions, i.e. high ambiguity and variance, is rare, though the extent of this complexity varies per task. The datasets evaluating WSD, SRL, and EL almost have a 1-to-1 mapping between lexical expressions and meanings, while coreference datasets have higher ambiguity and variance. This can be due to the following reasons: 1. Some of the coreference datasets deliberately focus on increasing ambiguity. 2. An inherent property of coreference seems to be high variance. Similarly, our dominance metrics (MODA/MODV and EMNLE/ELENM) demonstrate a strong bias in our datasets: typically, for any of the datasets, approximately 90% of the occurrences belong to the dominant class on average.

Concerning the concept-oriented tasks, Table 13 shows a notable difference in the complexity of the interaction between the proxies of datasets and resources.[29] Between 74 and 80% of the resource ambiguity per expression is not represented in the datasets, whereas this is the case for 60-64% of the resource variance per concept. This is an indication of strong semantic overfitting of the data to a small selection that is not representative for the full potential of expressions and meanings. Furthermore, we observe that this representativeness is relatively constant across concept datasets, which in part can be explained by the fact that the WSD and SRL datasets mainly stem from the same time period (Figure 17), and even from the same corpus (Hovy and Søgaard, 2015). One could argue that the data is correctly representing the natural complexity of a specific time period and genre but it does not challenge systems to be able to shift from one situation to another. We also note a temporal discrepancy between the concept- and instance-based datasets, with the instance-based systems being evaluated on more recent data.

28 The metrics and the analyses of the datasets can be found at `https://github.com/cltl/SemanticOverfitting`.

29 While computing RORA and RORV, we ignore cases with resource ambiguity and variance of 1.

Task	Dataset	MOA	MOV	MODA	MODV	EMNLE	ELENM
EL	AIDA test B	1.09	1.35	0.98	0.91	0.05	0.22
	WES2015	1.06	1.33	0.97	0.88	0.05	0.21
	MEANTIME	1.19	4.63	0.98	0.64	0.04	0.55
EnC	QuizBowl	1.59	1.80	0.92	0.74	0.13	0.46
EvC	ECB	1.61	3.87	0.89	0.61	0.19	0.65
	ECB+	2.09	3.40	0.85	0.66	0.27	0.57
	TAC KBP '15	4.97	1.22	0.69	0.94	0.47	0.12
WSD	SE2 AW	1.20	1.06	0.94	0.98	0.13	0.05
	SE3 task 1	1.21	1.05	0.94	0.98	0.13	0.04
	SE7 task 17	1.14	1.04	0.95	0.98	0.10	0.03
	SE10 task 17	1.25	1.06	0.93	0.98	0.13	0.05
	SE13 task 12	1.10	1.06	0.97	0.98	0.14	0.05
SRL	CoNLL04	1.20	1.00	0.96	1.00	0.09	0.00

Table 11: Observed ambiguity, variance and dominance.

Task	Dataset	ATR	ATRFU
EL	WES2015	1.92	0.53
EL	MEANTIME	1.51	0.51

Table 12: ATR and ATRFU values of the datasets.

Task	Dataset	RORA	RORV
WSD	SE2 AW	0.26	0.38
	SE3 task 1	0.23	0.37
	SE7 task 17	0.20	0.36
	SE10 task 17	0.25	0.40
	SE13 task 12	0.26	0.40
SRL	CoNLL04	0.63	1.00

Table 13: RORA and RORV values of the datasets.

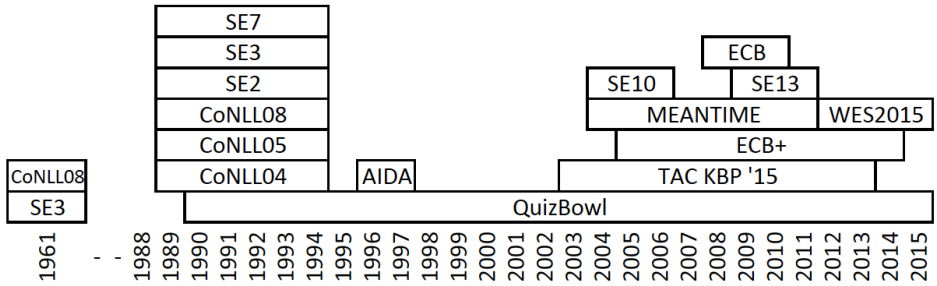


Figure 17: DTR values of the datasets

To understand further the time-bound interaction in our datasets, we study them together with time-bound resources. While our lexical resources and instance knowledge sources contain very little temporal information, we rely on query monitoring websites (Wikiviews and GoogleTrends) to get an indication of the usage of a meaning over time. In Table 12, we show the temporal popularity of entities among their candidates in our datasets according to our web sources.[30] We note a correspondence between the dominance of entities in datasets and

30 Due to the non-availability of information for the other tasks, we only analyze the temporal dominance for the EL task, even though the set of represented entities in DBpedia is not complete (as discussed in Section 3.6.3). In our analysis, we only consider ambiguous expressions that can denote more than one entity candidate. The candidates were obtained from the Wikipedia disambiguation pages. From Wikiviews, the month of the document creation time was used for the dominance information.

their frequency of usage at that time, which exposes a bias of existing datasets towards the most popular entities at the time of their creation.

Our analysis reveals that the existing disambiguation and reference datasets show a notable bias with respect to the aspects of ambiguity, variance, dominance, and time, thus exposing a strong semantic overfitting to a specific part of the world, while largely ignoring long-tail phenomena. Typically this is the part of the world that is best known within the context of a generation at a moment of time. This implies that our datasets have a strong bias towards meanings that are popular at that particular moment in time and do not represent the temporal relativity of this bias. Although our metrics provide us with a valuable set of insights into the evaluation datasets, complementary statistical measures should be introduced in the future to capture individual distinctions blurred by averaging over a dataset. These could measure the distribution of ambiguity and variance, their relation to dataset size, and outliers.

3.8 PROPOSAL FOR IMPROVING EVALUATION

The direct contribution of our work lies in metric-based evaluation of datasets and resources for systems, which helps interpreting their ability to cope with alterations of ambiguity, variance, dominance, and time.

Provided that a collection of multi-task annotated data is available at a central place, our metrics could be applied to output a dataset following certain criteria, e.g., a test set annotated with WSD and EL, whose ambiguity and variance are both between 1.2 and 1.4, and whose documents have been created in the 90s. The practical obstacle is the availability of input data, which can be addressed by the following (semi)automatic expansion method: 1. Collect annotated data and split the data according to time periods. 2. Collect annotated expressions from the data with their dominant meanings. 3. Retrieve new documents using unsupervised techniques in which these expressions occur with evidence for usage in other meanings than the dominant one in the existing datasets. Evidence can come from meta data, unsupervised clustering, and temporal and topical distance from annotated data. 4. Fix alternative meanings for all tokens in the new texts (one meaning-per-document), if necessary applying additional disambiguation tools. Add this data as silver data to the collection. 5. If necessary, re-annotate silver data manually or add annotations for other tasks.[31] 6. Spread documents over different time periods for both annotated gold data and silver data to obtain sufficient variation in time-bound contexts. Provided that this acquisition procedure is successful, selecting a dataset would require almost no effort, which enables creation of many, well-motivated datasets. Consequently, the dynamic nature of this process would challenge semantic overfitting.

In this thesis, we have not tested the method proposed here, since it turned out to be too hard/laborious to realize. Instead, we developed a more efficient, event-based Question Answering (QA) task, extensively elaborated on in Chapter 4. The metrics presented here are used to ensure that the task questions

31 Excessive labor could be avoided by prioritizing relevant expressions, e.g., according to the metrics.

exhibit maximum confusability in terms of ambiguity of words, surface forms, and events, as well as spatio-temporal ambiguity. In order to perform on this task with a good accuracy, the systems will be required to exhibit a deeper semantic understanding of the linguistic tail of the reference and disambiguation tasks we analyze in this chapter.

Time-aware evaluation datasets, originating from controlled methodologies as the ones described here, would allow the community to test understanding of language originating from different generations and communities, and a community's language usage in relation to different world contexts. It would also assess to what extent a semantic NLP system can adapt to language use from another time slice than the one trained on, with potentially new meanings and expressions, and certainly a different distribution of the expression-meaning relation. We believe this challenges semantic overfitting to one single part and time of the world, and will inspire systems to be more robust towards aspects of ambiguity, variance, and dominance, as well as their temporal dependency.

3.9 CONCLUSIONS

Disambiguation and reference systems tend to treat evaluation datasets as given, without questioning their representativeness, which makes it difficult to interpret efficacy improvements in terms of the individual contributions of algorithms and data. For system evaluations to provide generalizable insights, we must understand better the details of the disambiguation/reference task that a given dataset sets forth.

In this chapter we first described our preliminary analysis of a number of entity linking benchmark datasets with respect to an array of characteristics that can help us interpret the results of proposed entity linking systems. We measured that current evaluation datasets in the task of EL focus on head entities and expose very little ambiguity. Notably, the overlap of the concrete entities is fairly large between most datasets.

We then expanded this analysis to a general framework of metrics applicable to five disambiguation and reference tasks, including EL. For this purpose, we qualified and quantified the relation between expressions and meanings in the world for a generation sharing a language system, as well as the fluctuation in usage of expressions and meanings over time. We proposed the Semiotic Generation and Context Model, which captures the distribution changes in the semiotic relation given the context of the changing world. We applied it to address three key problems concerning semantic overfitting of datasets. We conceptualized and formalized generic metrics which evaluate aspects of datasets and provided evidence for their applicability on popular datasets with running text from five semantic tasks. We observed that existing disambiguation and reference datasets show a notable bias with respect to aspects of ambiguity, variance, dominance, and time, thus exposing a strong semantic overfitting to a very limited, and within that, popular part of the world. Finally, we proposed a time-based, metric-aware approach to create datasets in a systematic and semi-

automated way as well as an event-based QA task. Both approaches will result in datasets that would challenge semantic overfitting of disambiguation systems. The design and implementation of the latter proposal is discussed in detail in the upcoming chapter.

Answer to RQ2 This chapter investigated the second question of this thesis: *Are the current evaluation datasets and metrics representative for the long-tail cases?* Our model-based metrics revealed a representativeness bias for the task of EL, as well as four other disambiguation and reference tasks. Notably, existing datasets for any of the tasks in question exposed strong dominance, minimal ambiguity and variation, as well as strong temporal bias to specific periods in the past 50 years. Given that the underrepresented cases are often those of the tail, and that in chapter 2 we observed that the tail is far more challenging for systems, there is urgency for this representativeness bias to be addressed.

To do so, we start with the metric-driven task proposal for creation of event-based QA datasets detailed in section 3.8 of this chapter, and we develop it further, resulting in the first semantic NLP task that deliberately addresses the long-tail cases. The process of creation and the outcomes of this task provide evidence for the third research question of this thesis, RQ3: *How can we improve the evaluation on the long-tail cases?* This is covered in the next chapter.

4 IMPROVING THE EVALUATION BIAS ON THE LONG TAIL OF DISAMBIGUATION & REFERENCE

Although a name such as *Ronaldo* can have an infinite amount of references and in any world (real or imaginary) each *Ronaldo* is equally present, the previous chapter evidenced that our datasets usually make reference to only one *Ronaldo*. This lack of representativeness in our NLP tasks has big consequences for language models: they tend to capture the head phenomena in text without considering the context constraints and thus fail when dealing with less dominant world phenomena. As a result, there is little awareness of the full complexity of the task in relation to the contextual realities, given language as a system of expressions and the possible interpretations within context of time, location, community, and topic. Chapter 2 witnessed that this lack of awareness actually leads to low system performance on the tail. People, however, have no problem to handle local real-world situations that a text makes reference to.

We believe it is time to create a task that encourages systems to model the full complexity of disambiguation and reference by enriched context awareness.[1] For this purpose, we put forward a referential challenge for semantic NLP that reflects a higher degree of ambiguity and variation and captures a large range of small real-world phenomena. We held this referential quantification task at SemEval-2018 (Postma et al., 2018). With it, we address the third research question of this thesis (RQ3): *How can we improve the evaluation on the long-tail cases?*

The content of this chapter is based on the research published in the following four publications:

1. Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp (2016b). "Moving away from semantic overfitting in disambiguation datasets." In: *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. Austin, TX: Association for Computational Linguistics, pp. 17–21. URL: http://aclweb.org/anthology/W16-6004 Marten Postma and I have collaboratively created the majority of this paper. We worked together on all phases of the paper: motivation, literature review, methodology, proposal, and discussion, and only divided tasks within a phase.

2. Marten Postma, Filip Ilievski, and Piek Vossen (2018). "SemEval-2018 Task 5: Counting Events and Participants in the Long Tail." In: *Proceedings of the*

1 Please be aware that in this chapter the word "task" will be ambiguously used to refer to two different senses. As in the remainder of the thesis, "task" will refer to a generic NLP challenge (such as Entity Linking), whereas here it will often be used to denote a specific *evaluation task* instance, to which we will also simply refer as a "task".

12th International Workshop on Semantic Evaluation (SemEval-2018). New Orleans, LA, USA: Association for Computational Linguistics Marten Postma and I have collaboratively created the majority of this paper. The creation of the SemEval-2018 task 5 required the following steps: writing a task proposal, development of the data-to-text method, data collection, task design, task organization, paper writing, and presentation at the SemEval-2018 workshop. We worked together on all phases of this competition, and only divided tasks within a phase.

3. Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers (2018a). "Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* In this paper, I have contributed with the preparation of the underlying task data, creating an annotation environment, surveying existing structured datasets, and writing.

4. Piek Vossen, Marten Postma, and Filip Ilievski (2018b). "ReferenceNet: a semantic-pragmatic network for capturing reference relations." In: *Global Wordnet Conference 2018, Singapore* I have contributed to the writing and editing of various sections of this paper. Most of the paper was created by prof. Piek Vossen.

4.1 INTRODUCTION

The topic of this chapter is a "referential quantification" task that requires systems to establish the meaning, reference, and identity of events[2] and participants in news articles. By "referential quantification", we mean questions concerning the number of incidents of an event type (e.g., *How many killing incidents happened in 2016 in Columbus, MS?*) or participants in roles (e.g., *How many people were killed in 2016 in Columbus, MS?*), as opposed to factoid questions for specific properties of individual events and entities (e.g., *When was 2pac murdered?*). The questions are given with certain constraints on the location, time, participants, and event types, which requires understanding of the meaning of words mentioning these properties (e.g., Word Sense Disambiguation), but also adequately establishing the identity (e.g., reference and coreference) across mentions. The task thus represents both an intrinsic and application-based extrinsic evaluation, as systems are forced to resolve ambiguity of meaning and reference, as well as variation in reference in order to answer the questions.

Figure 18 shows an overview of our quantification task. We provide the participants with a set of questions and their corresponding news documents.[3] Systems are asked to distill event- and participant-based knowledge from the documents to answer the question. Systems submit both a numeric answer (3 events in Figure 18), and the corresponding events with their mentions found in the provided

2 By event, we denote a specific instance of an event, e.g., a killing incident happening at a specific location, time, and involving certain participants.

3 Question parsing is unnecessary, as questions are provided in a structured format.

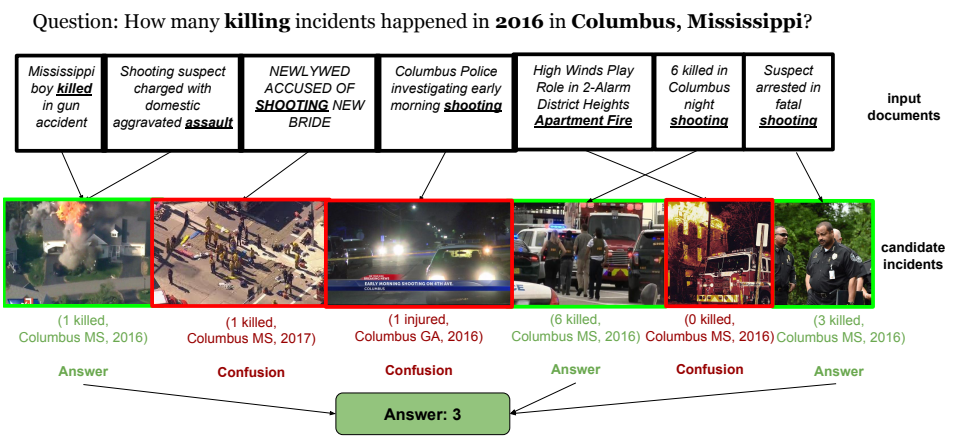


Figure 18: Task overview. Systems are provided with a question and a set of input documents. Their goal is then to find the documents that fit the question constraints and reason over them to provide an answer.

texts (e.g., the leftmost incident in Figure 18 is referred to by the coreferring mentions "killed" and "assault" found in two separate documents). Systems are evaluated on both the numeric answers as well as on the sets of coreferring mentions. Mentions are represented by tokens and offsets provided by the organizers.

The incidents and their corresponding news articles are obtained from structured databases, which greatly reduces the need for annotation and mainly requires validation instead. Given this data and using a metric-driven strategy, we created a task that further maximizes ambiguity and variation of the data in relation to the questions. This ambiguity and variation includes a substantial amount of low-frequent, local events and entities, reflecting a large variety of long-tail phenomena. As such, the task is not only highly ambiguous but can also not be tackled by relying on the most frequent and popular (head) interpretations.

We see the following contributions of our task:

1. To the best of our knowledge, we propose the first task that is deliberately designed to address large ambiguity of meaning and reference over a high number of infrequent, **long-tail** instances.

2. We introduce a methodology, called **data-to-text**, for creating large event-based tasks while avoiding a lot of annotation, since we base the task on structured data. The remaining annotation concerns targeted mentions given the structured data rather than full documents with open-ended interpretations.

3. We made all of our code to create the task available,[4] which may stimulate others to create more tasks and datasets that tackle long-tail phenomena

4 https://github.com/cltl/LongTailQATask

for other aspects of language processing, either within or outside of the SemEval competition.

4. This task provides insights into the strengths and weaknesses of semantic processing systems with respect to various long-tail phenomena. We expect that systems need to innovate by adjusting (deep) learning techniques to capture the referential complexity and knowledge sparseness, or by explicitly modeling aspects of events and entities to establish identity and reference.

4.2 MOTIVATION & TARGET COMMUNITIES

Expressions can have many different meanings and possibly an infinite number of references (consider, for instance, the word "number"). At the same time, variation in language is also large, as we can make reference to the same things in many ways. This makes the tasks of Word Sense Disambiguation, Entity Linking, and Event and Nominal Coreference extremely hard. It also makes it very difficult to create a task that represents the problem at its full scale. Any sample of text will reduce the problem to a small set of meanings and references, but also to meanings that are popular at that time excluding many unpopular ones from the distributional long tail. Given this Zipfian distribution, a task that is challenging with respect to ambiguity, reference, and variation, and that is representative for the long tail as well, needs to fit certain constraints.

Our task directly relates to the following communities in semantic processing:

1. disambiguation and reference;
2. reading comprehension and question answering.

4.2.1 *Disambiguation & reference*

Semantic NLP tasks are often limited in terms of the range of concepts and meanings that are covered. This is a necessary consequence of the annotation effort that is needed to create such tasks. Likewise, in Chapter 3, we observed that most well-known datasets for semantic tasks have an extremely low ambiguity and variation. Even in datasets that tried to increase the ambiguity and temporal diversity for the disambiguation and reference tasks, we still measured a notable bias with respect to ambiguity, variance, dominance, and time. Overall, tasks and their datasets show a strong semantic overfitting to the head of the distribution (the most popular part of the world) and are not representative for the diversity of the long tail.

Our task differs from existing ones in that: 1. we deliberately created a task with a high number of event instances per event type, many of which with similar properties, leading to high confusability; 2. we present an application-based task which requires to perform on a combination of intrinsic tasks such as ref-

erence, disambiguation, and spatial-temporal reasoning, that are usually tested separately in existing tasks.

4.2.2 *Reading Comprehension & Question Answering*

In several recent tasks, systems are asked to answer entity-based questions, typically by pointing to the correct segment or coreference chain in text, or by composing an answer by abstracting over multiple paragraphs/text pieces. These tasks are based on Wikipedia (SQuAD (Rajpurkar et al., 2016), WikiQA (Yang et al., 2015), QASent (Wang et al., 2007), WIKIREADING (Hewlett et al., 2016)) or on annotated individual documents (MS MARCO (Nguyen et al., 2016b), CNN, and DailyMail datasets (Hermann et al., 2015)). The questions in WikiQA (Yang et al., 2015) and QASent (Wang et al., 2007) are comprised of samples from validated user query logs, while the answers are annotated manually from automatically selected Wikipedia pages. WIKIREADING (Hewlett et al., 2016) is a recent large-scale dataset that is built based on the structured information from Wikidata (Vrandečić and Krötzsch, 2014) together with the unstructured information available in Wikipedia. Following a smart fully-automated data acquisition strategy, this dataset contains questions about 884 properties of 4.7 million instances. While these datasets require semantic text processing of the questions and the candidate answers, there is a finite set of answers, many of which represent popular interpretations from the world, as a direct consequence of using Wikipedia as an information source.

Weston et al. (2015) outlined 20 skill sets, such as causality, resolving time and location, and reasoning over world knowledge, that are needed to build an intelligent QA system that can deal with the linguistic phenomena of the long tail. These have been partially captured by the datasets MCTest (Richardson et al., 2013) and QuizBowl (Iyyer et al., 2014)), as well as the SemEval task on *Answer Selection in Community Question Answering* (Nakov et al., 2015, 2016).[5]

However, all these datasets avoid representing real-world referential ambiguity to its full extent by mainly asking questions that require knowledge about popular Wikipedia entities and/or text understanding of a single document.[6] As Wikipedia and Wikidata represent only a tiny and popular subset of all world events, the Wikipedia-based approaches could not be applied at all to create such task, thus signaling the need for a novel data acquisition approach to create an event-driven QA task for the long tail. Unlike existing work, our task deliberately addresses the referential ambiguity of the world beyond Wikipedia, by asking questions about long-tail events described in multiple documents. By doing so, we require deep processing of text and establishing identity and reference across documents.

5 The 2017 run can be found at `http://alt.qcri.org/semeval2017/task3/`.

6 e.g., the Quiz Bowl dataset deliberately focuses on domains with much training data and frequent answers, thus avoiding the long tail problem in reference.

4.2.3 *Moving away from semantic overfitting*

While the long tail has been partially captured in past tasks, none of these approaches has yet created a task that addresses the long tail explicitly and recognizes the full referential complexity of disambiguation. Since the field is highly competitive, a task for which it is necessary to perform well on the long tail phenomena would motivate systems that can deal with the long tail and towards systems that address the full complexity of the disambiguation task. To our knowledge, we propose the first QA task that deliberately addresses the problem of (co)reference to long tail instances, where the list of potential interpretations is enormous, largely ambiguous, and only relevant within a specific contextual setting.

4.3 TASK REQUIREMENTS

Our quantification task consists of questions like *How many killing incidents happened in 2016 in Columbus, MS?* on a dataset that maximizes confusability of meaning, reference, and identity. To guide the creation of such task, we defined five requirements that apply to the data for a single event type, e.g., *killing*.

Each event type should contain:

R1 Multiple event instances per event type, e.g., *the killing of Joe Doe* and *the killing of Joe Roe*.

R2 Multiple event mentions per event instance within the same document.

R3 Multiple documents with varying creation times that describe the same event. This requirement prevents strategies that rely only on the document creation times to achieve high performance on the task. In addition, documents that report on the same incident at very different time points tend to provide different, and sometimes even contradicting information, which challenges systems to reason over incomplete and sometimes wrong/outdated information. An example for such a case is a document that reports an updated (higher) number of casualties than the one reported earlier in another document.

R4 Event confusability by combining one or multiple confusion factors:

a) ambiguity of mentions across event types, e.g., *John Doe fires a gun*, and *John Doe fires a worker*.

b) variance of the mentions of an event instance, e.g., *John Doe kills Joe Roe*, and *John Doe murders Joe Roe*.

c) time, e.g., *killing A that happened in January 2013*, and *killing B in October 2016*.

d) participants, e.g., *killing A committed by John Doe*, and *killing B committed by Joe Roe*.

e) location, e.g., *killing A that happened in Columbus, MS*, and *killing B in Houston, TX*.

R5 Representation of non-dominant events and entities, i.e., instances that receive little media coverage. Hence, the entities would not be restricted to celebrities and the events are not widely discussed such as general elections, preventing them to be guessed from popularity/frequency priors.

4.4 METHODS FOR CREATING AN EVENT-BASED TASK

How can we create an event-based QA task that satisfies these requirements in an efficient manner? Event data so far is created with the text-to-data approach. We next discuss the limitations of this method and propose an alternative method for easier creation of large event datasets, which we call data-to-text.

4.4.1 *State of text-to-data datasets*

Event coreference annotations so far have been created using what we call a *text-to-data (T2D)* approach. In the T2D approach, annotators start from the text and first decide what phrases are labeled as event mentions after which different event mentions are related to each other through an event coreference relation. The coreference relations establish event identity across event mentions a posteriori by chaining event mentions that share coreference relations. Figure 19 gives a schematic overview for the T2D approach that indirectly constructs a referential representation from annotated mentions of events and participants.

We next discuss four main drawbacks of the T2D method: D1. size; D2. ambiguity and variation; D3. efficiency and scalability; D4. definition.

D1: SIZE Due to the complexity and labor-intensity of this T2D approach, only a limited amount of referential event data has been created so far (see Table 14). This table shows the number of documents in each dataset, the number of mentions of events, and the number of so-called coreference clusters (groups of mentions that refer to the same event). The final column indicates if the coreference clusters span across documents (cross-document coreference) or only within a single document (within-document coreference). We observe that the number of documents and mentions is small for both within- and cross-document relations: less than four thousand documents and less than forty thousand mentions in total (10 mentions per document on average). The ratios between mentions and clusters vary considerably, which is due to the different ways in which the datasets have been compiled: either a specific selection of the sentences (e.g., 1.8 sentences per article on average in ECB+) and/or event types (e.g., only violence or finance) were annotated, or all mentions in a full article.

D2: AMBIGUITY & VARIATION In chapter 3 we analyzed the referential annotations in a number of these datasets, revealing that they, despite efforts such

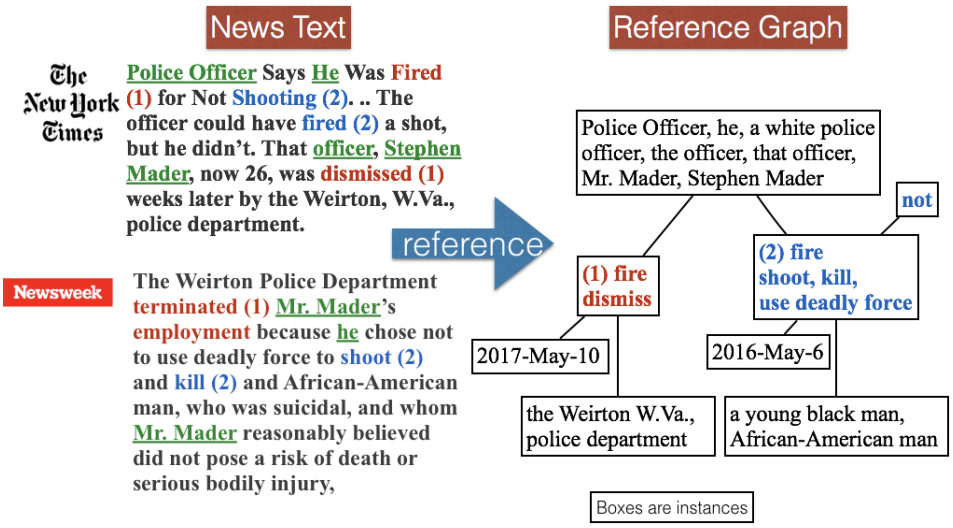


Figure 19: Overview of the T2D method: deriving a referential graph from mentions across different news text.

as the creation of ECB+, hardly reflect referential ambiguity and show very little variation. For example, ECB with 482 documents contains 8 news articles on one specific murder, but since there are no other murders in the dataset, searching for the word "murder" results in almost all mentions of that specific incident with high accuracy: one-form-one-referent and one-referent-one-form. Cybulska and Vossen (2014) demonstrated that the so-called lemma baseline to establish coreference relations[7] scores already very high in this dataset and is difficult to beat by state-of-the-art systems. From the perspective of a real-world situation and the many different ways in which events can be described and framed in language, these datasets are far too sparse and do not reflect true ambiguity and variation. Partly due to this lack of data and variation, automatic event coreference detection has made little progress over the years, especially across documents (Bejan and Harabagiu, 2010; Chen and Ji, 2009; Lee et al., 2012; Liu et al., 2014; Lu and Ng, 2016; Peng et al., 2016; Vossen and Cybulska, 2016).

D3: EFFICIENCY AND SCALABILITY All data listed in Table 14 are created according to the T2D approach: a selection of text is made and interpreted by annotators who add an annotation layer. Creating data following a T2D approach is expensive and labor-intense, as all mentions of events need to be cross-checked against all other mentions across documents for coreference relations. With the size of the data, the effort increases exponentially.

7 all occurrences of the same word, e.g., "murder", mention a single unique event and hence are coreferential.

Table 14: Event coreference corpora for English created by following the text-to-data method. For comparison, the last row presents statistics on The Gun Violence Corpus (GVC), whose development following our data-to-text method we detail later in this chapter.

Name	**Reference**	**nr. docs**	**nr mentions**	**mention/ docs.**	**nr clusters**	**mention/ cluster**	**cross doc.**
ACE2005	(Peng et al., 2016)	599	5268	8.79	4046	1.30	NO
KBP2015	(Mitamura et al., 2015a)	360	13113	36.43	2204	5.95	NO
OntoNotes	(Pradhan et al., 2007)	1187	3148	2.65	2983	1.06	NO
IC	(Hovy et al., 2013)	65	2665	41.00	1300	2.05	NO
EECB	(Lee et al., 2012)	482	2533	5.26	774	3.27	YES
ECB+	(Cybulska and Vossen, 2014)	982	6833	6.96	1958	3.49	YES
MEANTIME	(Minard et al., 2016)	120	2096	17.47	1717	1.22	YES
EER	(Hong et al., 2016)	79	636	8.05	75	8.48	YES
RED	(O'Gorman et al., 2016)	95	8731	91.91	2390	3.65	YES
Total		3874	36292	9.37	15057	2.41	
GVC	(Vossen et al., 2018a)	510	7298	14.31	1411	5.17	YES

D4: DEFINITION Guidelines and annotations tend to differ in criteria for deciding whether a mention represents an event and on the text span to be annotated as a mention. Regardless of the types of events annotated, these criteria are set a priori and they tend to vary depending on the specific task for which the data were annotated, e.g., semantic role detection (Kingsbury and Palmer, 2002), detecting temporal and causal event relations (Bethard et al., 2015; Boguraev et al., 2007; Caselli and Morante, 2016; Pustejovsky and Verhagen, 2009), or event coreference relations (Hovy et al., 2013). Such difficulties in defining events, event relations, and event coreference have led to the creation of the KBP2015 dataset (Mitamura et al., 2015a) in which a weaker definition of an event has been applied, so-called Event Nuggets, to ease the annotation and the task for establishing coreference relations. In the KBP2015 dataset, "attack", "shooting", and "murder" do not represent separate event instances, but are considered as mentions of the same underspecified event represented at a more coarse-grained level of granularity, so-called event-hoppers.

Overall, these drawbacks prohibit the controlled creation of a referential quantification task with a high volume and confusability. To create an event-based QA task following the requirements set in section 4.3, we would need to manually select a large quantity of relevant documents that correspond on the same and similar incidents, share temporal context, and contain resembling surface forms; and then extensively annotate coreference between the surface form mentions of events by a pairwise comparison between any two documents.

We attempted to automate this process in a similar way as the proposal in section 3.8, by: 1. indexing a million news articles from the SignalMedia corpus (Corney et al., 2016) in ElasticSearch; 2. querying for topical keywords to compile a set of documents about a topic[8]; 3. processing the data with the NewsReader pipeline (Vossen et al., 2016) to generate semantic representation in the form of a queryable knowledge graph; 4. query this graph in a smart way to generate questions with high confusability. Unfortunately, this automatic method yielded low-quality results, whereas the manual method proposed above is far too laborious to apply it to realistically generate a task with high volume and ambiguity.

For these reasons, in the next section we propose a new method that starts from registered events that are given a priori when annotating event references in texts so that we only need to compare mentions across relevant documents, knowing in advance what events are being covered in these documents.

4.4.2 *From data to text*

The research on event coreference faces a data bottleneck because it is both too difficult and too costly to gather sufficient data following the traditional T2D method. We therefore propose a novel *structured-data-to-text* (D2T) methodology,

8 Example query to retrieve flood disasters: `http://news.fii800.lod.labs.vu.nl/news?q=flood%20disaster&match=conjunct&to=2015-09-30T00:00:00Z&in=content&media=News&size=50&from=2015-09-01T00:00:00Z`.

based on the notions **microworlds** and **reference texts**. *Microworlds* are structured representations of referents related to specific world events (e.g., human calamities or economic events). *Reference texts* are documents reporting on this data, e.g., news articles, blogs, and Wikipedia pages. In the D2T method, we start from some event registry that has been created by people a priori by hand and is publicly available as structured data. From these registries, we derive microworld representations of the unique event instances, their participants, location, and date as a referential graph, as shown in Figure 20. Assuming that reference texts mainly refer to the corresponding microworld and not to other events and participants, we can establish the referential relation relatively easily and partially automatically.

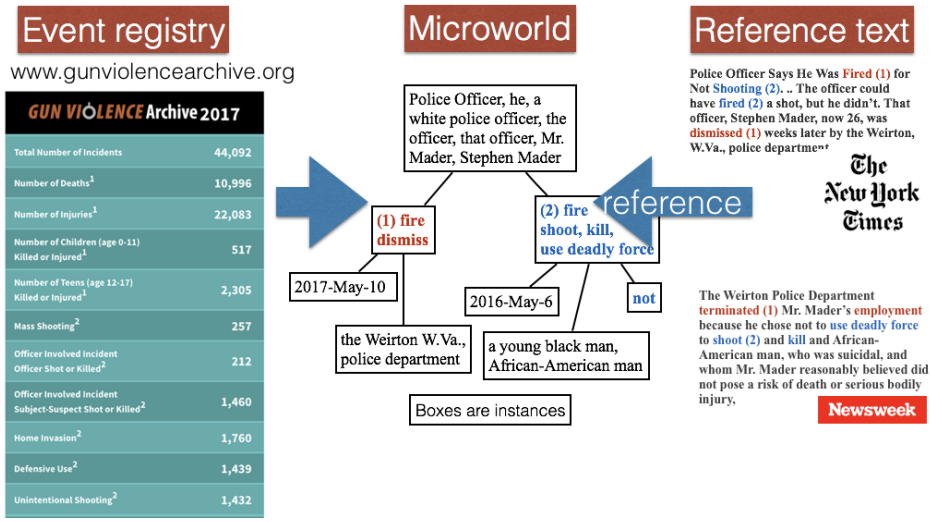


Figure 20: Overview of the D2T method: representing structured event data first as microworlds and secondly pairing it with reference texts.

Now we explain how the D2T method overcomes the main drawbacks of T2D as captured in the points D1-D4 above. By combining microworlds for similar but different events with their paired reference texts, we increase the referential ambiguity for systems that need to reconstruct the microworld from the texts, hence approximating the complexity of reference relations in reality across large volumes of text. In this manner, D2T notably improves on the drawbacks *D1: size* and *D2: ambiguity*. Later in this chapter we also provide empirical evidence for this improvement. The volumes of event data that can be extracted with D2T (Table 15) are orders of magnitude higher than those of existing corpora created with T2D (Table 14): 836,000 against less than 4,000 documents. Similarly, the confusability in our task exceeds that of T2D.[9] By collecting news from different sources on the same or similar events, we better approximate the true variation in

9 For each gold event instance there are close to a hundred that share some, but not all, event properties with it.

making reference from different perspectives (*D2: variance*). We provide evidence for the variance captured in our task in section 4.7. Furthermore, the fact that the data on events from which we start has been created from the perspective of general human interest (e.g., gun violence incident reports) avoids the never-ending discussion on what establishes an event in text (*D4: definition*). More practically, the D2T method is much less labor-intensive than T2D (*D3, efficiency*), because a rich and consistent set of event properties and links to its supporting documents are provided within a microworld by the original data - we discuss this in section 4.7. Finally, since the underlying data is often created manually, its quality is very high.

The D2T method can thus create a lot of ambiguous data in an automatic way, while still retaining high quality. Next, we discuss the data properties that are desired by this method, and several resources that fit most of these desiderata.

Desiderata The promise of the D2T approach comes at the price of more strict requirements about the underlying data. While the T2D method only expects a set of text documents, the expectations of D2T are far more challenging. Namely, our method operates best on resources with:

1. *links between structured data and reporting texts*
2. *disambiguated/unique and consistently defined* events and event properties following Linked Data principles
3. *open, available* data
4. *high volume*, since more data typically exhibits higher referential ambiguity and variation.

If all four desiderata are fulfilled, the conversion of the data to microworlds and reference texts is a matter of writing data manipulation scripts. In practice, resource properties are often not ideal, thus requiring some additional work - however, the amount of annotation or retrieval needed is far lower/incomparable to the exhaustive annotation processes in T2D.

Resource availability Table 15 provides description of several public resources that satisfy most of the desiderata. The resources register event incidents with rich properties such as participants, location, and incident time, and they provide pointers to one or more reference texts. The number of events and documents is usually high, for instance there are ~9K incidents in the Railways Archive (RA), and ~231K incidents in the Gun Violence Archive (GVA).[10]

10 In addition, data with similar properties can be obtained from Wikipedia and structured databases such as Wikidata, YAGO2, and DBpedia with little effort. This can either be done through direct extraction, or through smart querying of the data (Elbassuoni et al., 2010; Hewlett et al., 2016; Knuth et al., 2015). For example, a simple query on Wikidata for event instances belonging to certain event classes (i.e. explosion, crime, natural disaster, accident, sport, election), already yields over 70k events with structured data (type of event, location and time) that can form the basis for creating microworlds. Many of these events can be traced back to Wikipedia pages, that describe these events in textual form. Such Wikipedia pages often include further links to news articles as references to substantiate the information given. By using Wikipedia as the glue between the structured microworld data and the reference texts, one can obtain a reliable mapping of texts with framings and representations of the referential events.

Table 15: Potential event data for extracting microworlds and reference texts. Numbers marked with '*' are estimates.

Name	Topic	Structured data	Nr docs	Nr incidents	From year	To year	Locations	Reference texts
ASN incident database[a]	aircraft safety occurrences	fatalities, locations, time, other domain data	32K	21K	1919	2017	world	news, reports, social media
ASN Wikibase[b]	aircraft safety occurrences	fatalities, locations, time, other domain data	310K	207K	1905	2017	world	news, reports, social media
Fire Incident Reports (FR)[c]	fire disasters	publishing time and location	1K	1K	2004	present	USA	reports
Global nonviolent action DB[d]	social justice/protests	incident location and time	*6K	1K	1955	present	world	various
Gun Violence Archive (GVA)[e]	gun violence	fatalities, locations, time, participant roles, weapon information	*462K	231K	2012	present	USA	news
Legible news[f]	science, sports, business, economics, law, crime, disasters, accidents, ...	/	*20K	*15K	2014	present	world	news
Railways Archive (RA)[g]	railway accidents	casualties, locations, time, vehicle operators	5K	9K	1803	present	UK, Ireland	news
TOTAL			***836K**	***485K**				

a https://aviation-safety.net/database/
b https://aviation-safety.net/wikibase/
c https://www.firerescue1.com/incident-reports/
d https://nvdatabase.swarthmore.edu/
e http://gunviolencearchive.org/reports/
f http://legiblenews.com
g http://www.railwaysarchive.co.uk/eventlisting.php

Following the D2T method, we successfully obtained data from these resources with extreme ambiguity and variation, while maintaining the identity and reference relations and without having to annotate large quantities of texts word-by-word. Concretely, we gathered over ten thousand news articles and over five thousand incidents from GVA and FR (Fire Incident Reports), which were used as a basis for the referential quantification task described in this chapter.

4.5 DATA & RESOURCES

In this Section, we present our data sources, obtained with the D2T method, as well as an example document. We also discuss considerations of licensing and availability.

4.5.1 *Structured data*

The majority of the source texts in our referential quantification task are sampled from structured databases that contain supportive news sources about gun violence incidents. While these texts already contain enough confusability with respect to the aspects defined in Section 4.3, we add confusion through leveraging structured data from two other domains: fire incidents and business.

As a direct consequence of using these databases and our exploitation strategy, we are able to satisfy all requirements we set in Section 4.3. These databases contain many event instances per event type (R1), multiple event mentions in the same document per event instance (R2), mentions of an event instance across multiple documents with a wide spread of publishing times (R3), represent non-dominant events and entities (R5), and contain rich annotation of event properties that allows us to create high confusability (R4, see Section 4.6.3 for our methodology).

For a large portion of the information in the structured databases, we manually validated that this information could be found in the supportive news sources, and excluded the documents for which this was not the case. For the remaining documents, we performed automatic tests to filter out low-quality entries.

GUN VIOLENCE The gun violence data is collected from the standard reports provided by the *Gun Violence Archive (GVA)* website.[11] Each incident contains information about: 1. its **location** 2. its **time** 3. how many people were **killed** 4. how many people were **injured** 5. its **participants**. Participant information includes: a) the **role**, i.e., victim or suspect b) the **name** c) the **age** 6. the **news articles** describing this incident. Table 16 provides a more detailed overview of the information available in the GVA.

To prevent systems from cheating (by using the structured data directly), the set of incidents and news articles is extended with news articles from the Signal-1M Dataset (Corney et al., 2016) and from the Web, that also stem from the gun violence domain, but are not found in the GVA.

11 http://gunviolencearchive.org/reports/

Event Property	Granularity	Example value
Location	Address	Central Avenue
	City	Waynesboro
	State	Mississippi
Incident time	Day	14-3-2017
	Month	3-2017
	Year	2017
Participant	First name	John
	Last name	Smith
	Full name	John Smith

Table 16: Overview of the GVA incident properties of location, time, and participant.

OTHER DOMAINS For the fire incidents domain, we make use of the *FireRescue1* reports,[12] which describe the following information about 417 incidents: 1. their **location** as a surface form 2. their **reporting time** 3. one **free text summary** describing the incidents 4. **no** information about **participants**. Based on this information, we manually annotated the incident time and mapped the location to its representation in Wikipedia.

We further carefully selected a small number of news articles from the business domain from The Signal-1M Dataset, by querying for "people fired". Since these documents were not semantically annotated with respect to event information, we manually annotated this data with the same kind of information as the other databases: incident location, time, and information on the affected participants.

4.5.2 *Example document*

For each document, we provide its **title**, **content (tokenized)**, and **creation time**, for example:
Title: *$70K reward in deadly shooting near N. Philadelphia school*
Content: *A $70,000 reward is being offered for information in a quadruple shooting near a Roman Catholic school ...*
DCT: *2017-4-5*

4.5.3 *Licensing & availability*

The news documents in our task are published on a very diverse set of (commercial) websites. Due to this diversity, there is no easy mechanism to check

12 https://www.firerescue1.com/incident-reports/

their licenses individually. Instead, we overcome potential licensing issues by distributing the data under the Fair Use policy.[13] [14]

During the SemEval-2018 period, but also afterwards, systems can easily test their submissions via our competition on CodaLab.[15]

4.6 TASK DESIGN

For every incident in the task, we have fine-grained structured data with respect to its event type, location, time, and participants, and unstructured data in the form of the news sources that report on it. In this Section, we explain how we exploited this data in order to create the task. We present our three subtasks and the question template after which we outline the question creation. Finally, we explain how we divided the data into trial and test sets and provide some statistics about the data. For detailed information about the task, e.g., about the question and answer representation, we refer to the CodaLab website of the task.

4.6.1 *Subtasks*

The task contains two event-based subtasks and one entity-based subtask.

Subtask 1 (S1): Find the single event that answers the question e.g., *Which killing incident happened in Wilmington, CA in June 2014?* The main challenge is not to determine how many incidents satisfy the question, but to identify the documents that describe the single answer incident.

Subtask 2 (S2): Find all events (if any) that answer the question, e.g., *How many killing incidents happened in Wilmington, CA in June 2014?* This subtask differs from S1 in that the system now also has to determine the number of answer incidents, which makes this subtask harder. To make it more realistic, we also include questions with zero as an answer.

Of course, we use different questions in S1 compared to those in S2 (see below for details on how these questions were created), to fit the different assumptions underlying these two subtasks. Namely, we ensure that all questions in S1 are about a single incident, and we use questions with a wide range of number of incidents as an answer in S2.

Subtask 3 (S3): Find all participant-role relations that answer the question e.g., *How many people were killed in Wilmington, CA with the last name Smith?* The goal is to determine the number of entities that satisfy the question. The system not only needs to identify the relevant incidents, but also to reason over identities and roles of the participants.

4.6.2 *Question template*

Questions in each subtask consist of an event type and two event properties.

13 Fair use policy in USA: `https://goo.gl/hXiEKL`
14 Fair use policy in EU: `https://goo.gl/s8V5Zs`
15 `https://competitions.codalab.org/competitions/17285`

EVENT TYPE We consider four event types in this task described through their representation in WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 2003). Each question is constrained by exactly one event type.

event type	description	meanings
killing	at least one person is killed	wn30:killing.n.02 wn30:kill.v.01 fn17:Killing
injuring	at least one person is injured	wn30:injure.v.01 wn30:injured.a.01 fn17:Cause_harm fn17:Experience-_bodily_harm
fire_burning	the event of something burning	wn30:fire.n.01 fn17:Fire_burning
job_firing	terminated employment	wn30:displace.v.03 fn17:Firing

Table 17: Description of the event types. The meanings column lists meanings that best describe the event type. It contains both FrameNet 1.7 frames (prefixed by *fn17*) and WordNet 3.0 synsets (prefixed by *wn30*).

EVENT PROPERTIES For each event property in our task (time, location, participants), we distinguish between three levels of granularity (see Table 16). In addition, we make a distinction between the surface form and the meaning of an event property value. For example, the surface form *Wilmington* can denote several meanings: the Wilmington cities in the states of California, North Carolina, and Delaware. When composing questions, for time and location we take the semantic (meaning) level, while for participants we use the surface form of their names. This is because the vast majority of the participants in our task are long-tail instances which have no semantic representation in a structured knowledge base.

4.6.3 *Question creation*

Our question creation strategy consists of three consecutive phases: question composition, generation of answer and confusion sets, and question scoring. These steps are common for both the event-based subtasks (S1 and S2) and the entity-based subtask S3.

1. **Question composition** We compose questions based on the template described in Section 4.6.2. This entails:
   a) choice of a subtask
   b) choice of an event type, e.g., *killing*
   c) choice of two event properties (e.g., *time and location*) with their corresponding granularities (e.g., *month and city*) and concrete values (e.g., *June 2014 and Wilmington, CA*).

   This step generates a vast amount of potential questions (hundreds of thousands) in a data-driven way, i.e., we select the event type and properties per question purely based on the combinations we find in our data. Example questions are:

   Which killing event happened in June 2014 in Wilmington, CA?
   (subtask S1)

   How many killing events happened in June 2014 in Wilmington, CA?
   (subtask S2)

   How many people were killed in June 2014 in Wilmington, CA?
   (subtask S3)

2. **Answer and confusion sets generation** For each generated question, we define a set of answer and confusion incidents with their corresponding documents. **Answer** incidents are the ones which entirely fit the question parameters, e.g., *all killing incidents that occur in June 2014 and in the city of Wilmington, CA*. **Confusion** incidents fit some, but not all, values of the question parameters, i.e., they differ with respect to an event type or property (e.g., *all fire incidents in June 2014 in Wilmington, CA*; or *all killings in June 2014, but not in Wilmington, CA*; or *all killings in Wilmington, CA, but not in June 2014*).

3. **Question scoring** The generated questions with their corresponding answers and confusion are next scored with respect to several metrics that measure their complexity. These metrics are inspired by the methodology and the analysis in chapter 3. The per-question scores allow us to detect and remove the "easy" ones, and keep those that:
   a) have a high number of answer incidents (only applicable to S2 and S3)
   b) have a high number of confusion incidents
   c) have a high average number of answer and confusion documents, i.e., news sources describing the answer and the confusion incidents correspondingly
   d) have a high temporal spread with respect to the publishing dates reporting on each incident from the answer and confusion incidents

e) have a high ambiguity with respect to the surface forms of an event property value in a granularity level (e.g., we would favor *Wilmington*, since it is a city in at least three US states in our task data).

	S	#Qs	Avg answer	Avg # answer docs
trial	1	424	1.00	1.68
	2	469	4.22	7.68
	3	585	5.48	5.47
test	1	1032	1.00	1.60
	2	997	3.79	6.64
	3	2456	3.66	3.74

Table 18: General statistics about trial and test data. For each subtask (*S*), we show the number of questions (*#Qs*), the average answer (*Avg answer*), and the average number of answer documents (*Avg # answer docs*).

4.6.4 *Data partitioning*

We divided the overall task data into two partitions: **trial** and **test** data. In practice, we separated these two data partitions by reserving one year of news documents (2017) from our task for the trial data, while using all the other data as test data.

The trial data stems from the gun violence domain, whereas the test data also contains data from the fire incidents and business domain. A subset of the trial and test data has been annotated for event coreference. Table 18 presents the most important statistics of the trial and test data.

We made an effort to make the trial data representative for the test data with respect to the main aspects of our task: its referential complexity, high confusability, and long-tail instances. Despite the fact that the trial data contains less questions than the test data, Table 18 shows that it is similar to the test data with respect to the core properties, meaning that the trial data can be used as training data.

4.7 MENTION ANNOTATION

The execution of the task design described in the previous section (4.6) suffices to evaluate the numerical answer and the set of answer documents provided by a system. In addition, it guarantees high confusability and low dominance of the underlying textual data with respect to the posed questions.

In order to evaluate the third aspect provided by the systems, namely the mention-level annotation of the incidents, we perform mention annotation on a

subset of the task data. Following the D2T method proposed in this chapter, the annotation of mentions can then be seen merely as a task of marking evidence for the incident and its characteristics in the supporting text documents. The results of the mention annotation as described in this section were released after SemEval-2018 as the Gun Violence Corpus (GVC) on top of the data created for mention evaluation. This section describes the details of this annotation and the development of GVC.

4.7.1 *Annotation task and guidelines*

The annotation of a mention process involved three basic steps:

- Annotating the *event type* of every mention that refers to a gun violence incident in the structured data;
- Annotating the *victim(s)* involved in the mention referring to a shooting in the structured data;
- Annotating every mention related to gun violence but NOT referring to the incident in the structured data (*other incidents or generic mentions*).

Based on these annotations, we can infer coreference relations: in case that two or more mentions have the same annotations (event type and victims) AND they both relate to the same incident ID in the structured data, we can infer that these mentions are coreferential. Since the annotation is done simultaneously on all documents that describe a given incident, this strategy automatically infers cross-document coreference relation.

To further capture the referential complexity and diversity of event descriptions in text, we designed an event schema that captures subevent relations in addition to the above incident references, see Figure 21. The main event ("the gun incident") is basically a container that can be split into several more fine-grained events that stand in some implication relation to each other. Following (Cybulska and Vossen, 2015), we denominate this main event container *Bag of events*. In this case the bag of events consists of five events: *Firing a gun*, *Hitting someone*, or *Missing someone*. An event of *Hitting someone* can lead to *Injuring* and in some cases *Death*. Apart from these events, many articles also contain references to gun violence in a more general way or not related to the structured data. These have been labeled *Generic* and *Other*.

We annotated all mentions denoting but also implying one of the predefined event classes. For example, a *funeral*, an *autopsy*, or the process of *grieving* imply that someone died. A *shooter* and *killer* imply respectively the event types *Firing a gun* and again *Death* in the context of this domain. Besides annotating verbal and nominal expressions, we also annotated mentions of other parts of speech (including adjectives and adverbs), idioms, multi-word units, and collocations. In principle, we annotated the minimal span of a mention, usually the head, unless this would result in a meaningless annotation, e.g., we would annotate *critical condition* as a multi-word unit instead of just the head *condition*.

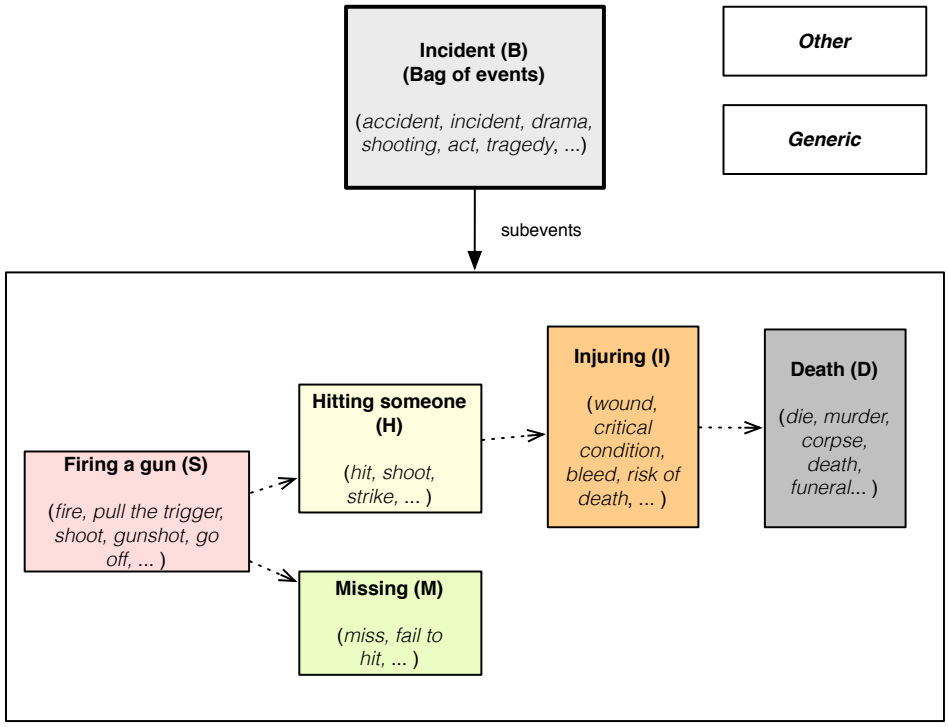


Figure 21: The event scheme used for the annotation of gun violence events.

Additional specification of the annotation decisions, such as: how we handled negation, the irrealis, ellipsis, phrasal verbs, and various cases of implicit event mentions, can be found in the full guidelines.[16]

4.7.2 *Annotation environment*

To the best of our knowledge, there is no tool that starts from structured event data to annotate event mentions and event coreference relations. We therefore built our own environment for annotating events in reference texts that are related to structured data on an incident.

The goal of the tool is to allow annotators to find evidence in all the reference texts for the event properties in the structured data. To support this goal, the tool reads the structured event data and presents the event properties, e.g., time, location, and participants, in a table. Annotators mark the event mentions, select the participants involved and select the type of event. The annotators only need to annotate the mentions of the predefined schema and not all other types of events.

16 The full guidelines are available at `https://goo.gl/Yj1Hra`.

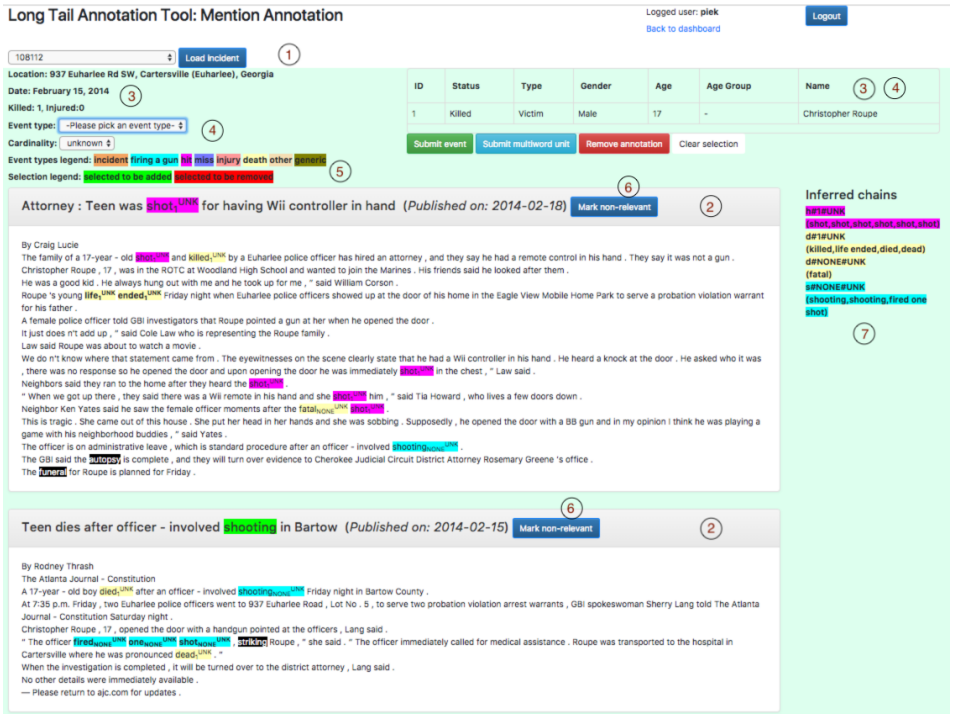


Figure 22: Annotation environment for annotating mentions in Reference Texts related to structured data.

By applying this strategy to all mentions within and across Reference Texts of an incident, we establish coreference and identity across the mentions. Notably, it is not needed to annotate coreference explicitly. Instead, the coreference chains are inferred by the annotation environment, based on the combination of two factors of the individual mention annotations: event type and participants, with the a priori event data.

In addition, we have built in lexical support for the annotators, based on the set of already annotated event mentions. Reference text mentions which have been frequently annotated in other texts but not in the current one, are visually prompted for annotation. The annotators can then decide whether to accept this suggestion.

Figure 22 provides a screenshot of the mention annotation environment when the incident *108112* is loaded by the user *piek*. The incident selection menu is marked with (1) in the Figure. The selected incident is supported by two reference texts, rendered in the middle of the screen (marked with (2)). Annotators can select one or multiple mentions from this area for annotation. The top panel contains the structured data about the current incident (marked with (3)), followed by menus and a table for annotations of properties for the selected mention (4). Mentions in colors have already been annotated by this user, and the

event type is signaled by the color. The color scheme is explained in detail in the legend (5). Moreover, inversely colored mentions (e.g., "funeral" and "autopsy" in Figure 22) are the ones proposed by the tool to be annotated additionally. Annotators can also discard individual documents with the 'Mark non-relevant' button (6). Finally, the area on the right displays the coreferential chains that the tool has inferred so far about the current incident (marked with (7)).

The source code of the annotation software is available on Github.[17]

4.7.3 *Annotation process*

Two linguistic students were hired to perform the annotations. After completing the training phase, which resulted in some simplifications of the guidelines, the students started with the mention annotation. In six weeks, the students annotated the 510 documents that are part of the corpus. In addition, 25 documents were selected in order to compute the inter-annotator agreement (IAA). The first annotator annotated 432 event mentions in this set, whereas the second one annotated 457 event mentions. The annotators provided the same annotation in 350 cases, resulting in a Cohen's kappa coefficient (Cohen, 1960) of 0.72.[18] According to Landis and Koch (1977), a score between 0.61 and 0.80 is considered *substantial*, from which we conclude that there was high agreement between the annotators. For comparison, ECB+ (Cybulska and Vossen, 2014) reported a Cohen's kappa coefficient of 0.68 for a similar size and agreement analysis to ours. ECB+ annotators only had to consider 2 incidents per topic with about 10 articles per incident and 1.8 sentences on average per article, whereas in our case, 510 documents need to be annotated for a few hundred incidents. In terms of speed, one annotator averaged 5 minutes per document, whereas the other took 4 minutes to annotate one document on average.

As argued in section 4.4, our method scales only linearly instead of exponentially, unlike in T2D. Namely, to include documents that report on a new incident, one does not need to compare their mentions to all other incidents, since the structured data already guarantees they are not coreferential. In Table 14, we report statistics on the size of our corpus. Although our corpus annotated with mentions is currently smaller than existing datasets, the speed and the linear scalability of our method provide a promise that its size can increase up to the limit posed by the original structured data sources.

4.7.4 *Corpus description*

The Gun Violence Corpus (GVC),[19] resulting from this annotation, contains 7,298 mentions, referring to 241 incidents. In total, 510 documents contain at least one mention. Table 19 presents the annotation frequency for each event type.

17 `https://github.com/cltl/LongTailAnnotation`

18 We observed that the first annotator was more consistently following the guidelines. Hence, we resolved the remaining disagreements by using her annotation in the final corpus.

19 The corpus can be downloaded at: `https://github.com/cltl/GunViolenceCorpus`

event type	annotation frequency
Death	2,206
Firing a gun	1,622
Hitting	1,122
Bag of events	755
Injuring	726
Other	596
Generic	270
Missing	2

Table 19: Mention frequency of each event type.

Most mentions in our Gun Violence Corpus refer to the event types *Death* and *Firing a gun*, respectively. In addition, about 4% of all mentions (i.e., 270 mentions), refer to generic uses of shooting and killings. Finally, it is not uncommon that the text refers to other incidents than the main incident of the article, which happens in about 8% of all mentions (i.e., 596). This means that systems cannot fully rely on a one-incident-per-document heuristic to detect coreference chains.

event type	most common expressions
Death	dead (305) died (285) killed (283)
Firing a gun	shooting (680) gunshot (247)
	went off (72)
Hitting	shot (801) shooting (83) struck (46)
Bag of events	shooting (247) incident (164) it (88)
Injuring	wound (175) injured (75) injuries (68)
Other	shot (105) shooting (70) killed (47)
Generic	accident (57) shooting (13) tragedy (11)
Missing	surgery (1) missed (1)

Table 20: Most common expressions used for event types

Table 20 presents the most used expressions for each event type. As presented in this Table, the most common expressions that are used to refer to event types are covered well in resources such as WordNet. For example, the most common expressions for the event type *Death* can be detected by correctly identifying the WordNet synsets *kill.v.01* (cause to die; put to death, usually intentionally or knowingly) and *killing.n.02* (the act of terminating a life). However, this is not the case for all expressions in the GVC. For example, expressions like *mourn* and *autopsy* that refer to the event type *Death* show that manual and automatic

annotators cannot fully rely on resources to detect all event types correctly, but that additional reasoning is needed.

In (Vossen et al., 2018b), we analyze the referential potential of this corpus further and we propose to use it to build a **ReferenceNet** on top of WordNet, to capture the pragmatics of language use beyond semantics, relating to information sharing, relevance, salience, and framing. In table 21 we provide the resulting ReferenceNet for the event types considered in the GVC corpus, on 20 processed incidents with 38 documents in total. We observe that the event implications follow from very different expressions. For example, *Death* can be concluded forward from *fatal shot* or backward from *autopsy*. Especially words making reference to the complete incident show a lot of variation, reflecting different judgments and appraisals. The specifics of this ReferenceNet proposal and the automatic approaches for deriving it through the D2T and the T2D methods fall out of the scope of this thesis; we refer the reader to the original paper for an extensive description.

Table 21: ReferenceNet at the event type level, derived from manual annotation for 38 news documents reporting on 20 gun violence incidents. The annotation resulted in 138 event instances and 874 mentions in 38 documents. In total, 77 different lemmas were used to make reference to these events. Given these annotations, we can abstract from the instances and group lemmas that make reference to the same *type* of event. Note that the total number of mentions and lemmas is higher, as the same word, e.g., *shooting* may occur in multiple reference sets.

Event type	Nr. Variants	Nr. Mentions	ReferenceSets
Bag of events	27	229	accident:39, incident:34, it:34, this:17, murder:15, hunting:14, reckless:14, tragedy:9, happen:8, felony:7, manslaughter:5, what:5, homicide:4, shooting:4, assault:3, case:2, endanger:2, endangerment:2, that:2, violence:2, crime:1, event:1, mistake:1, situation:1
Firing a gun	21	148	shooting:48, fire:25, discharge:16, go:12, shot:9, pull:7, gunman:6, gun:5, gunshot:4, firing:3, shoot:2, turn:2 , accidental:1, act:1, action:1, at:1, handle:1, it:1, return:1, shootout:1, shotgun:1,
Hitting	11	196	shot:131, discharge:17, shooting:17, strike:16, hit:4, blast:3, victim:3, striking:2, gunshot:1, into:1, turn:1
Injuring	16	73	wound:36, surgery:13, treat:5, injure:3, stable:3, injurious:2, send:2, bodily:1, critical:1, hit:1, hospitalize:1, hurt:1, injury:1, put:1, stabilize:1, unresponsive:1
Death	16	246	death:60, die:52, dead:45, kill:34, fatal:13, lose:9, fatally:7, loss:7, autopsy:6, body:4, take:3, homicide:2, claim:1, deadly:1, life:1, murder:1
Total	114	1043	

4.8 EVALUATION

This section describes the evaluation criteria in our referential quantification task and the baselines we compare against.

4.8.1 *Criteria*

Evaluation is performed on three levels: incident-level, document-level, and mention-level.

1. **The incident-level evaluation** compares the numeric answer provided by the system to the gold answer for each of the questions. The comparison is done twofold: by exact matching and by Root Mean Square Error (RMSE) for difference scoring. The scores per subtask are then averaged over all questions to compute a single incident-level evaluation score.
2. **The document-level evaluation** compares the set of answer documents between the system and the gold standard, resulting in a value for the customary metrics of Precision, Recall, and F1 per question. The scores per subtask are then averaged over all questions to compute a single document-level evaluation score.
3. **The mention-level evaluation** is a cross-document event coreference evaluation. Mention-level evaluation is only done for questions with the event types *killing* or *injuring*. We apply the customary metrics to score the event coreference: BCUB (Bagga and Baldwin, 1998), BLANC (Recasens and Hovy, 2011), entity-based CEAF (CEAF_E) and mention-based CEAF (CEAF_M) (Luo, 2005), and MUC (Vilain et al., 1995). The final F1-score is the average of the F1-scores of the individual metrics. The annotation of mentions was explained in Section 4.7.

4.8.2 *Baselines*

To stimulate participation in general and to encourage approaches beyond surface form or majority class strategies, we implemented a baseline to infer incidents per subtask and a baseline for mention annotation.[20]

INCIDENT INFERENCE BASELINE This baseline uses surface forms based on the question components to find the answer documents. We only consider documents that contain the label of the event type or at least one of its WordNet synonyms. The labels of locations and participants are queried directly in the document (e.g., if the location requested is the *US state of Texas*, then we only consider documents that contain the surface form *Texas*, and similarly for participants such as *John*). The temporal constraint is handled differently: we only consider documents whose publishing date falls within the time period requested in the question.

20 The code of the baselines can be found here: `https://goo.gl/MwSqBj`.

For subtask 1, this baseline assumes that all documents that fit the created constraints are referring to the same incident. If there is no such document, then the baseline does not answer the question (because S1 always has at least one supporting document). For subtask 2, we assume that none of the documents are coreferential. Hence, if 10 documents match the constraints, we infer that there are also 10 corresponding incidents. No baseline was implemented for subtask 3.

MENTION ANNOTATION BASELINE We annotate mentions of events of type *killing* and *injuring*, when these surface forms or their synonyms in WordNet are found as tokens in a document. We assume that all mentions of the same event type within a document are coreferential, whereas all mentions found in different documents are not.

4.9 PARTICIPANTS

Next, we describe the systems that took part in SemEval-2018 task 5. We refer to the individual system papers for further information.

NEWSREADER (NWR) (Vossen, 2018) consists of three steps: 1. the event mentions in the input documents are represented as Event-Centric Knowledge Graphs (ECKGs); 2. the ECKGs of all documents are compared to each other to decide which documents refer to the same incident, resulting in an incident-document index; 3. the constraints of each question (its event type, time, participant names, and location) are matched with the stored ECKGs, resulting in a number of incidents and source documents for each question.

NAI-SEA (Liu and Li, 2018) consists of three components: 1. extraction of basic information on time, location, and participants with regular expressions, named entity recognition, and term matching; 2. event classification with an SVM classifier; 3. document similarity by applying a classifier to detect similar documents. In terms of resources, NAI-SEA combines the training data with data on American cities, counties, and states.

FEUP (Abreu and Oliveira, 2018) developed an experimental system to extract entities from news articles for the sake of Question & Answering. For this main task, the team proposed a supervised learning approach to enable the recognition of two different types of entities: Locations (e.g., *Birmingham*) and Participants (e.g., *John List*). They have also studied the use of distance-based algorithms (using Levenshtein distance and Q-grams) for the detection of documents' closeness based on entities extracted.

ID-DE (Mirza et al., 2018) created KOI (Knowledge of Incidents), a system that builds a knowledge graph of incidents, given news articles as input. The required steps include: 1. document preprocessing using various semantic NLP tasks such as Word Sense Disambiguation, Named Entity Recognition,

R	Team	s2_inc_acc norm	s2_inc_acc (% of Qs answered)	s2_inc rmse
1	FEUP	**26.38**	26.38 (100.0%)	6.13
2	*NWR	21.87	21.87 (100.0%)	43.96
3	Baseline	18.25	18.25 (100.0%)	8.50
4	NAI-SEA	17.35	17.35 (100.0%)	20.59
5	ID-DE	13.74	20.36 (67.5%)	6.15

Table 22: For subtask 2, we report the normalized incident-level accuracy (*s2_inc_acc norm*), the accuracy on the answered questions only (*s2_inc_acc*), and the RMSE value (*s2_inc rmse*). Systems are ordered by their rank (*R*).

Temporal expression recognition, and Semantic Role Labeling; 2. incident extraction and document clustering based on the output of step 1; 3. ontology construction to capture the knowledge model from incidents and documents which makes it possible to run semantic queries on the ontology to answer the questions, by using the SPARQL query language.[21]

4.10 RESULTS

Before we report the system results, we introduce a few clarifications regarding the result tables:

1. For the incident- and document-level evaluation, we report both the performance with respect to the subset of questions answered and a **normalized score**, which indicates the performance on all questions of a subtask. If a submission provides answers for all questions, the normalized score will be the same as the non-normalized score.
2. Contrary to the other metrics, a lower **RMSE** value indicates better system performance. In addition, the RMSE scores have not been normalized since it is not reasonable to set a default value for non-answered questions.
3. **The mention-level evaluation** was the same across all three subtasks. For this reason, results are only reported once (see section 4.10.3).
4. The teams whose member **co-organized SemEval-2018 task 5** are marked explicitly with an asterisk in the results.

4.10.1 *Incident-level evaluation*

The incident-level evaluation assesses whether the system provided the right numeric answer to a question. The results of this evaluation are given in the

21 https://www.w3.org/TR/rdf-sparql-query/

R	Team	s3_inc_acc norm	s3_inc_acc (% of Qs answered)	s3_inc rmse
1	FEUP	**30.42**	30.42 (100.0%)	478.71
2	*NWR	21.05	21.05 (100.0%)	296.45
3	NAI-SEA	20.20	20.2 (100.0%)	13.45
4	ID-DE	12.87	19.32 (66.61%)	7.87

Table 23: For subtask 3, we report the normalized incident-level accuracy (*s3_inc_acc norm*), the accuracy on the answered questions only (*s3_inc_acc*), and the RMSE value (*s3_inc rmse*). Systems are ordered by their rank (*R*).

Event type	Subtask	#Qs	FEUP	ID-DE	NAI-SEA	*NWR	Baseline
fire_burning	S2	79	40.51	-	31.65	39.24	**49.37**
	S3	0	-	-	-	-	-
injuring	S2	543	**21.92**	^13.44	14.36	21.73	17.68
	S3	1502	**30.49**	^8.39	16.78	23.17	-
job_firing	S2	4	0.0	-	25.0	25.0	**50.0**
	S3	26	**30.77**	-	26.92	15.38	-
killing	S2	371	**30.19**	^17.25	18.6	18.33	12.13
	S3	928	**30.28**	^20.47	25.54	17.78	-

Table 24: For subtask 2 (*S2*) and subtask 3 (*S3*), we report the incident-level accuracy and the number of questions (#*Qs*) per event type. The best result per event type for a subtask is marked in bold. '^' indicates that the accuracy is normalized for the number of answered questions, in cases where a system answered a subset of all questions.

Tables 22 and 23, for the subtasks 2 and 3 correspondingly.[22] On both subtasks, the order of the participating systems is identical, team *FEUP* having the highest score.

These tables also show the RMSE values, which measure the proximity between the system and the gold answer, punishing cases where the absolute difference between them is large. While for subtask 2 the system with the lowest error rate corresponds to the system with the highest accuracy, this is different for subtask 3. *ID-DE*, ranked last in terms of accuracy, has the lowest RMSE. This means that although their answers were not exactly correct, they were on average much closer to the correct answer than those of the other systems. This is more notable in subtask 3 since here the range of answers is larger than in subtask 2 (the maximum answer in subtask 3 is 171).

22 Incident-level evaluation was not performed for subtask 1, because per definition, its answer is always 1.

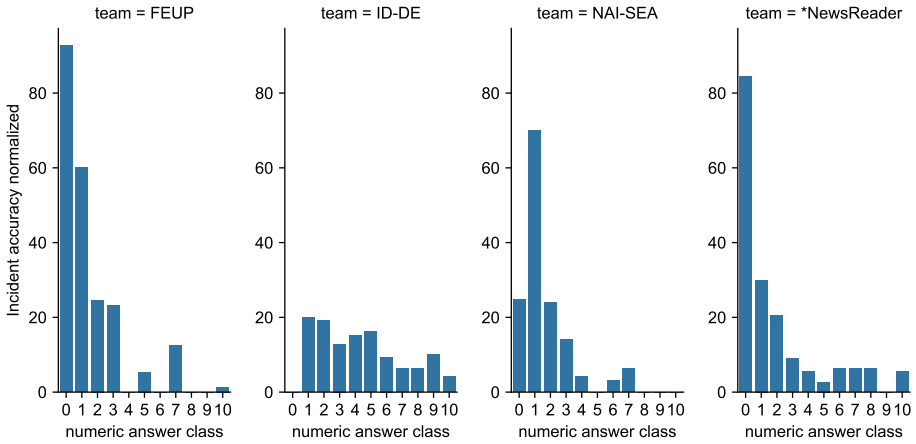


Figure 23: Incident-level accuracy of all systems per numeric answer class for subtask 2. The class *10* represents all answers of 10 or higher.

We performed additional analysis to compare the performance of systems per subtype and per numeric answer class. Table 24 shows that the system *FEUP* is not only superior in terms of incident-level accuracy overall, but this is also mirrored for most of the event types, especially those corresponding to the gun violence domain. On the other hand, Figure 23 shows the accuracy distribution of each system per answer class. Notably, for most systems the accuracy is highest for the questions with answer 0 or 1, and gradually declines for higher answers, forming a Zipfian-like distribution. The exception here is the team *ID-DE*, whose accuracy is almost uniformly spread across the various answer classes.

4.10.2 *Document-level evaluation*

The intent behind document-level evaluation is to assess the ability of systems to distinguish between answer and non-answer documents. The tables 25, 26, and 27 present the F1-scores for the subtasks 1, 2, and 3, respectively. Curiously, the system ranking is very different and almost opposite compared to the incident-level rankings, with the system *NAI-SEA* being the one with the highest F1-score. This can be explained by the multifaceted nature of this task, in which different systems may optimize for different goals.

Next, we investigated the F1-scores of systems per event property pair. As shown in Table 28, the best-performing system consistently has the highest performance over all pairs of event properties.

R	Team	s1_doc_f1 norm	s1_doc_f1 (% of Qs answered)
1	NAI-SEA	**78.33**	78.33 (100.0%)
2	ID-DE	36.67	82.99 (44.19%)
3	FEUP	24.65	24.65 (100.0%)
4	*NWR	23.82	46.2 (51.55%)
5	Baseline	11.09	67.33 (16.47%)

Table 25: For subtask 1, we report the normalized document-level F1 (*s1_doc_f1 norm*) and the accuracy on the answered questions only (*s1_doc_f1*). Systems are ordered by their rank (*R*).

R	Team	s2_doc_f1 norm	s2_doc_f1 (% of Qs answered)
1	NAI-SEA	**50.52**	50.52 (100.0%)
2	ID-DE	37.24	55.16 (67.5%)
3	*NWR	36.91	36.91 (100.0%)
4	FEUP	30.51	30.51 (100.0%)
5	Baseline	26.38	26.38 (100.0%)

Table 26: For subtask 2, we report the normalized document-level F1 (*s2_doc_f1 norm*) and the accuracy on the answered questions only (*s2_doc_f1*). Systems are ordered by their rank (*R*).

R	Team	s3_doc_f1 norm	s3_doc_f1 (% of Qs answered)
1	NAI-SEA	**63.59**	63.59 (100.0%)
2	ID-DE	46.33	69.56 (66.61%)
3	*NWR	26.84	26.84 (100.0%)
4	FEUP	26.79	26.79 (100.0%)

Table 27: For subtask 3, we report the normalized document-level F1 (*s3_doc_f1 norm*) and the accuracy on the answered questions only (*s3_doc_f1*). Systems are ordered by their rank (*R*).

Property pair	Task	#Qs	FEUP	ID-DE	NAI-SEA	*NWR	Baseline
location	S1	594	23.06	^26.64	**82.91**	^26.22	^8.71
& time	S2	680	30.95	^41.81	**49.99**	39.22	28.61
	S3	1335	26.4	^41.55	**63.27**	36.15	-
participant	S1	140	13.48	^43.86	**70.22**	^11.83	^9.76
& location	S2	49	14.66	^21.26	**50.41**	13.53	10.02
	S3	301	14.2	^44.28	**62.38**	6.65	-
participant	S1	298	33.06	^53.28	**73.01**	^24.65	^16.47
& time	S2	268	32.27	^28.55	**51.87**	35.34	23.71
	S3	820	32.06	^54.88	**64.56**	19.09	-

Table 28: Document-level F1-score and number of questions (#*Qs*) for each subtask (*S1, S2, and S3*) and event property pair as given in the task questions. The best result per property pair for a subtask is marked in bold. '^' indicates that the F1-score is normalized for the number of answered questions, in cases where a system answered a subset of all questions.

R	Team	BCUB	BLANC	CEAF_E	CEAF_M	MUC	AVG
1	ID-DE	**44.61%**	**31.59%**	37.45%	**47.23%**	**53.12%**	**42.8%**
2	*NWR	37.28%	28.11%	**42.15%**	46.16%	46.29%	40.0%
3	Baseline	6.14%	0.89%	13.3%	8.45%	3.59%	6.47%

Table 29: Results for mention-level evaluation, scored with the customary event coreference metrics: BCUB (Bagga and Baldwin, 1998), BLANC (Recasens and Hovy, 2011), entity-based CEAF (CEAF_E) and mention-based CEAF (CEAF_M) (Luo, 2005), and MUC (Vilain et al., 1995). The individual scores are averaged in a single number (*AVG*), which is used to rank (*R*) the systems.

4.10.3 *Mention-level evaluation*

Table 29 shows the event coreference results for the participating systems: *ID-DE* and *NewsReader (NWR)*, as well as our baseline. The columns present the F1-score for the metrics BCUB, BLANC, CEAF_E, CEAF_M, and MUC. The final column indicates the mean F1-score over these five metrics, which is used to rank the participants. The Table shows that the system *ID-DE* has a slightly better event coreference score on average over all metrics than the second-ranked system, *NewsReader*.

4.11 DISCUSSION

All four teams submitted a result for all three subtasks, and two teams participated in the mention-level evaluation. We observed that the ranking of systems

differs dramatically per evaluation level. The best document-level performance was obtained by the system *NAI-SEA*, which is understandable considering that this system's architecture relies on document retrieval. For providing numerical answers on the number of incidents, the system *FEUP* has the highest accuracy while the system *ID-DE* the lowest. Interestingly, the latter has much lower RMSE overall. In other words, while *ID-DE* is not able to provide the exact correct answer as often as *FEUP*, it is typically much closer to it. We also observed that while the performance of most systems overfits to low numerical answers (0 or 1), the performance of *ID-DE* is fairly constant across the classes of numerical answers. Given the multifaceted nature of this task, it is not surprising that systems chose different techniques and optimized for different goals.

Notably, although the systems are able to retrieve many of the answer documents, the highest accuracy of counting events or participants is 30%. Similarly, the mention coreference F1-scores of systems are between 28 and 53% per metric, which are much lower than the scores reported on the ECB+ corpus (cf. by Cybulska and Vossen (2014)). These observations can be explained with the complexity of this task, especially in terms of its ambiguity, dominance, and variance. The obtained results suggest that further research is necessary in order to develop complete and robust models that can natively deal with the challenge of counting referential units within sparse and ambiguous textual data.

The referential quantification task covered in this chapter does not evaluate the identity of long-tail entities directly, but only indirectly through a downstream QA task on counting events and participants. Nevertheless, we observe that entities of type *Location* and *Person* play an important role in all four participating systems, which suggests that the long-tail entity identity is very relevant for counting events and participants in local and ambiguous data. We investigate the task of establishing identity of the long-tail entities directly on this data in chapter 6.

4.12 CONCLUSIONS

In this chapter we addressed the question *RQ3* of this thesis: *How can we improve the evaluation on the long-tail cases?*, aiming to fill a gap in the representativeness of evaluation datasets for semantic NLP tasks, described and analyzed in chapter 3.

The traditional approach of text-to-data cannot be efficiently applied to create large-scale evaluation data with high quality, ambiguity, and variation. For this purpose, we first propose a novel method called data-to-text, that enables easier, scalable, and more flexible creation of large evaluation datasets, based on existing links between structured data and news documents that provide evidence for it.

We applied the data-to-text method to create a referential quantification task of counting events and participants in local news articles with high ambiguity. We organized this task as Task 5 at the SemEval-2018 competition. The complexity of this referential quantification task challenges systems to establish the mean-

ing, reference, and identity across documents. SemEval-2018 Task 5 consists of two subtasks of counting events, and one subtask of counting event participants in their corresponding roles. We evaluated system performance with a set of metrics, on three levels: incident-, document-, and mention-level.

For the mention evaluation, we created the Gun Violence Corpus: the first event coreference corpus developed following the data-to-text method. We show that we achieve high agreement and annotation speed, and report statistics of the resulting corpus. For future works, we aim to compare our annotation process to traditional annotation using text-to-data tools such as CAT (Lenzi et al., 2012) to annotate the same documents used in this study.

We described the approaches and presented the results of four participating systems, as well as two baseline algorithms. All four teams submitted a result for all three subtasks, and two teams participated in the mention-level evaluation. We observed that the ranking of systems differs dramatically per evaluation level. Given the multifaceted nature of this task, it is not surprising that different systems optimized for different goals. Although the systems are able to retrieve many of the answer documents, the highest accuracy of counting events or participants is 30%. This suggests that further research is necessary in order to develop complete and robust models that can natively deal with the challenge of counting referential units within sparse and ambiguous textual data.

Out-of-competition participation is enabled by the CodaLab platform, where this task was hosted.

Answer to RQ3 This chapter addressed the issue of semantic overfitting in disambiguation and reference datasets. Earlier in this thesis we observed that existing disambiguation datasets expose lack of representativeness and bias towards the head interpretation, while largely ignoring the rich set of long-tail phenomena. Semantic NLP systems are discouraged to consider the full referential complexity of the disambiguation task, since the main incentive lies in modelling the head phenomena.

As a response, we created the first long-tail task on semantic processing and held it at SemEval-2018. The task satisfied several requirements regarding the ambiguity, dominance, and variance of the concepts and instances it covers. As such, the extreme confusability of this task challenged systems to develop strategies that can address tail cases rather than relying on frequent observations, and to combine evidence from various disambiguation and reference tasks. The scores on this referential quantification task by all systems are fairly low (the highest accuracy being 30%), which confirms our expectations that the long-tail problem is difficult, under-addressed, and multifaceted.

Our referential quantification task has hence fulfilled its goal: to be representative for the tail cases. Did we follow the correct approach to create a long-tail task? Arbitrary sampling of the natural distribution of texts can hardly be expected to be representative for the distributional tail. We thus designed each step in the task creation process to be suitable for a long-tail task: 1. we listed explicit requirements for an ideal long-tail task 2. we developed a suitable method to obtain high-quality data on the tail with minimal annotation effort 3. we selected

data that is suitable to create a long-tail task with high confusability 4. following our requirements, we designed a task on top of this data that maximizes confusability. A potential downside to this task is its artificiality in terms of sampling, as we designed the data and the task to maximize confusability without preserving the natural distribution. We see this as a necessary trade-off in order to create a task that deliberately targets the evaluation bias on the tail, resulting in large confusability and requiring an extremely low annotation effort.

Given its properties and the system results obtained so far, we expect that this task will continue to attract attention in the upcoming years. In addition, other long-tail tasks can be created in the future by exploiting links between structured and unstructured data on long-tail instances, following the data-to-text method proposed here.

The referential quantification task covered in this chapter evaluates the identity of long-tail entities only indirectly through a downstream task on counting events and participants. Intuitively, we expect that the long-tail entities play an important role in this task as these are typically participants and locations in the events. In addition, the identity of people should be established in order to count event participants in subtask S3. This intuition is supported by the fact that entity detection is a key component of all four participating systems in this task. We investigate the task of establishing identity of the long-tail entities on this data directly in chapter 6.

So far, we have provided a distinction between the head and the tail cases, we have measured biases in the representativeness of current evaluation datasets with respect to the tail, and we addressed these biases by creating a new task that deliberately evaluates long-tail cases. Besides the bias in evaluation, however, similar biases can be found in the background knowledge sources that are used by NLP systems. We address these biases in the following chapter, thus attending to RQ4 of this thesis.

5 ENABLING ACCESS TO KNOWLEDGE ON THE LONG-TAIL ENTITIES BEYOND DBPEDIA

Analogous biases to those observed in evaluation can be found in the sources of knowledge that are leveraged by EL tools. Namely, current EL systems (DBpedia Spotlight (Daiber et al., 2013), Babelfy (Moro et al., 2014), NERD tools (Rizzo and Troncy, 2012)) rely on DBpedia or another Wikipedia-based knowledge source, and thus suffer from limited coverage (Kittur et al., 2009). These handful of knowledge sources contain general-purpose world knowledge, making them very suitable for linking of head (E1) entities from any domain, but unsatisfactory for interpreting tail (E2) and NIL (E3) entities. In this chapter we hence pose the fourth research question of this thesis (RQ4): *How can the knowledge on the long-tail entities be accessed and enriched beyond DBpedia?* We make an effort to improve the access to knowledge on these long-tail entities by leveraging the vast amount and diversity of knowledge found in the Linked Open Data cloud.

The content of this chapter is based on the research published in the following four publications:

1. Filip Ilievski, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2015). "LOTUS: Linked Open Text UnleaShed." In: *Proceedings of the Consuming Linked Data (COLD) workshop* In this paper, I am the main contributor to the design, implementation, and evaluation of LOTUS.

2. Filip Ilievski, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2016b). "LOTUS: Adaptive Text Search for Big Linked Data." In: *European Semantic Web Conference (ESWC) 2016*. Springer International Publishing, pp. 470–485 In this paper, I am the main contributor to the design, implementation, and evaluation of LOTUS.

3. Wouter Beek, Laurens Rietveld, Filip Ilievski, and Stefan Schlobach (2017). "LOD Lab: Scalable Linked Data Processing." In: *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering*. Lecture notes of summer school. Springer International Publishing, pp. 124–155 In this paper, I have contributed by designing and building the LOD Lab text search engine, LOTUS.

4. Wouter Beek, Filip Ilievski, Jeremy Debattista, Stefan Schlobach, and Jan Wielemaker (2018). "Literally better: Analyzing and improving the quality of literals." In: *Semantic Web Journal (SWJ)* 9.1, pp. 131–150. DOI: 10.3233/SW-170288. URL: https://doi.org/10.3233/SW-170288 In this paper, I

have contributed by performing the analysis of the language tags of RDF literals.

5.1 INTRODUCTION

The question that we seek to answer in this chapter is how to find existing structured representations of long-tail entities. As we have seen in chapter 3, current EL evaluation datasets include a small number of tail (E2) and NIL (E3) entities. This situation is partially due to the scarce amount of knowledge on long-tail entities in the background knowledge source (usually DBpedia) that a system links the surface forms to. Indeed, most long-tail entities are either scarcely or not at all represented in Wikipedia, and consequently its derivatives, like DBpedia. Note that Wikipedia[1] and Wikidata[2] contain a notability clause, containing the general guideline that a topic needs to have "significant coverage in reliable sources that are independent of the subject", in order to be worthwhile of representation in Wikipedia/Wikidata. Hence, a long-tail entity linking to DBpedia would be largely nonsensical, as most long-tail entities would have either no representation in DBpedia (NIL/E3 entities), or a representation with questionable value for linking (E2 entities). How can we then envision a version of the EL task that is suitable for linking to long-tail entities?

Linked Open Data sources, such as those found in the Linked Open Data Cloud[3], hold a promise for relatively easy access to a wealth of information. DBpedia, while being a central hub in the LOD cloud, is only one of the many thousands of knowledge sources found in it. By design, each of the datasets in the LOD cloud collection pertains to a universal RDF format, where facts are expressed in triples or quads, consisting of: a subject, a predicate, an object, and optionally, a graph identifier. This uniformity across the various knowledge sources that constitute the LOD cloud holds a potential for a broader definition of the entity linking task, beyond DBpedia.

Linking to entities beyond DBpedia is unfortunately hard to imagine at this point. This is because it has always been difficult to *find* resources on the Semantic Web: there is no centralized query service and the support for natural language access is limited. Simultaneous querying of a large number of knowledge sources is facilitated through services such as the LOD Laundromat (Beek et al., 2014), which provide a centralized copy of the LOD cloud. LOD Laundromat greatly improves the findability of the knowledge in the LOD cloud, but its querying options prior to this work are limited to structured queries. Namely, it is still not possible to find resources based on textual descriptions associated to that resource in the LOD Laundromat. In cases where a text index for Semantic Web data has been built, such an index has usually been restricted to a single knowledge base, as is apparent from the indices built within EL systems. The lack of a global entry point to resources through a flexible text index is a seri-

1 https://en.wikipedia.org/wiki/Wikipedia:Notability
2 https://www.wikidata.org/wiki/Wikidata:Notability
3 http://lod-cloud.net/

ous obstacle for Linked Data consumption in general, and an extra challenge for users from outside of the Semantic Web community that wish to use semantic resources. Enabling a natural language access to the LOD Cloud is a hard requirement for its usage in NLP applications, like entity linking systems.

This chapter describes the process of creation of such a centralized text index over the LOD cloud. In Section 5.8 we demonstrate the potential of this enhanced infrastructure to access entities that are found in various knowledge sources in the LOD cloud.

We introduce LOTUS: Linked Open Text UnleaShed, a central text-based entry point to a large subset of today's LOD Cloud. Centralized text search on the LOD Cloud is not new as Sindice[4] and LOD Cache[5] show. However, LOTUS differs from these previous approaches in three ways: 1. its scale (its index is about 100 times bigger than Sindice's was) 2. the adaptability of its algorithms and data collection 3. its integration with a novel Linked Data publishing and consumption ecosystem that does not depend on IRI dereferenceability.

LOTUS indexes every natural language literal from the LOD Laundromat data collection, a cached copy of a large subset of today's LOD Cloud spanning tens of billions of ground statements. The task of resource retrieval is a two-part process consisting of matching and ranking. Since there is no single combination of matching and ranking that is optimal for every use case, LOTUS enables users to customize the resource retrieval to their needs by choosing from a variety of matching and ranking algorithms. LOTUS is not a semantic search engine intended for end users, but a framework for researchers and developers in which semantic search engines can be developed and evaluated.

As we discussed in this thesis, entity linking on the tail entities is not a resolved challenge, and addressing it requires access to (more) knowledge on these rare entities beyond what DBpedia provides us. An adaptive linguistic entry point to the LOD Cloud, such as LOTUS, holds the potential to address the limitation of the currently used knowledge sources, and facilitate a Web-of-data-wide search and linking of entities. Furthermore, LOTUS might inspire new ideas for other research areas. For instance, as it provides a link between text and documents in the LOD cloud, Information Retrieval over the LOD cloud becomes an interesting option.

The remainder of this chapter is structured as follows. In Section 5.2, we detail the problem of performing linguistic search on the LOD Cloud and we formalize it by defining an array of requirements. Section 5.3 presents relevant previous work on Semantic Web text search. LOTUS is a first step up to an accessible disambiguation system over the LOD cloud - we discuss its position within a broader vision for evaluation and application infrastructure centered around the LOD Laundromat in Section 5.4. Section 5.5 describes the LOTUS framework through its model, approach, and initial collection of matching and ranking algorithms. The implementation of LOTUS is reported in Section 5.6. This Section also explains how to query LOTUS and provides concrete examples. Scalability

4 https://web.archive.org/web/20140104093943/http://www.sindice.com/, discontinued in 2014.
5 http://lod.openlinksw.com/

tests and demonstrations of LOTUS' flexible retrieval are discussed in Section 5.7. Several typical usage scenarios of LOTUS are presented and tested in Section 5.8. These use cases evaluate the potential of LOTUS for increasing the recall of the entity linking task and providing access to more knowledge on the tail entities, beyond what is found on DBpedia. We conclude by considering the key strengths, limitations, and future plans for LOTUS in Section 5.9. Here we also reflect on the fourth research question of this thesis.

5.2 PROBLEM DESCRIPTION

In this section, we detail the requirements for a global text-based entry point to linked open data and the strengths and weaknesses of current findability strategies with respect to these requirements.

5.2.1 *Requirements*

The Semantic Web currently relies on four main strategies to find relevant resources: datadumps, IRI dereferencing, Linked Data Fragments (LDF), and SPARQL, but these are not particularly suited to global text-based search.[6]

Since it is difficult to memorize IRIs and structured querying requires prior knowledge of how the data is organized, text-based search for resources is an important requirement for findability on the Semantic Web (Req1: Text-based). Furthermore, we also require text-based search to be resilient with respect to minor variations such as typos or spelling variations (Req2: Resilience). An important principle of the Semantic Web is that anybody can say anything about any topic (AAA). The findability correlate of this principle is not implemented by existing approaches: only what an authority says or links to explicitly can be easily found. We formulate the correlate requirement as "anybody should be able to find anything that has been said about any topic" (Req3: AFAA).

Decentralized data publishing makes text-based search over multiple data sources difficult. Not all sources have high availability (especially a problem for SPARQL) and results from many different sources have to be integrated on the client side (especially a challenge for IRI dereferencing and Linked Data Fragments). Hence, we set requirements on availability (Req4: Availability) and scalability (Req5: Scalability) for a text-based search service. While LDF provides better serviceability for singular endpoints than the other three approaches, it is still cumbersome to search for resources across many endpoints (Req6: Serviceability). Existing Semantic Web access approaches do not implement IR-level search facilities. In Section 5.3 some systems will be discussed that built some of this functionality on top of standard access methods. However, these systems focus on a single algorithm to work in each and every case. This may be suitable for an end-user search engine but not for a framework in which search engines

6 For the convenience of readers that might lack the needed background knowledge about these strategies, we provide an explanation in section 5.2.2.

are built and evaluated, bringing us to our last requirement of customizability (Req7: customizeability). Below we iterate our requirements:

REQ1. TEXT-BASED Resource-denoting IRIs should be findable based on text-based queries that match (parts of) literals that are asserted about that IRI, possibly by multiple sources.

REQ2. RESILIENCE Text-based search should be resilient against typos and small variations in spelling (i.e., string similarity and fuzzy matching in addition to substring matching).

REQ3. AFAA Authoritative and non-authoritative statements should both be findable.

REQ4. AVAILABILITY Finding resources should not depend on the availability of all the original resource-publishing sources.

REQ5. SCALABILITY Resources should be searchable on a Web scale, spanning tens of billions of ground statements over hundreds of thousands of datasets.

REQ6. SERVICEABILITY The search API must be freely available for humans (Web UI) and machines (REST) alike.

REQ7. CUSTOMIZABILITY Search results should be ranked according to a customizable collection of rankings to support a wide range of use cases.

5.2.2 *Current state-of-the-art*

DATADUMPS implement a rather simple way of finding resource-denoting terms: one must know the exact Web address of a datadump in order to download and extract resource-denoting IRIs. This means that search is neither text-based (Req1) nor resilient (Req2). Extraction has to be performed manually by the user resulting in low serviceability (Req6). Datadumps do not link explicitly to assertions about the same resource that are published by other sources (Req3).

IRIS DEREFERENCE to a set of statements in which that IRI appears in the subject position or, optionally, object position. Which statements belong to the dereference result set is decided by the authority of that IRI, i.e., the person or organization that pays for the domain that appears in the IRI's authority component. Non-authoritative statements about the same IRI cannot be found. Non-authoritative statements can only be found accidentally by navigating the interconnected graph of dereferencing IRIs. As blank nodes do not dereference significant parts of the graph cannot be traversed. This is not only a theoretical problem as 7% of all RDF terms are blank nodes (Hogan et al., 2012). In practice, this means that non-authoritative

assertions are generally not findable (Req3). Since only IRIs can be dereferenced, text-based access to the Semantic Web cannot be gained at all through dereferencing (Req1). Thus, it is not possible to find a resource-denoting IRI based on words that appear in RDF literals to which it is (directly) related, or based on keywords that bear close similarity to (some of the) literals to which the IRI is related (Req2).

LINKED DATA FRAGMENTS (Verborgh et al., 2014) (LDF) significantly increases the serviceability level for single-source Linked Data retrieval (Req6) by returning metadata descriptions about the returned data. E.g., by implementing pagination LDF allows all statements about a given resource to be extracted without enforcing arbitrary limits. This is not guaranteed by IRI dereferencing or SPARQL (both of which lack pagination). An extension to LDF (Van Herwegen et al., 2015) adds efficient substring matching, implementing a limited form of text-based search (Req1). Due to the reduced hardware requirements for running an LDF endpoint it is reasonable to assume that LDF endpoints will have higher availability than SPARQL endpoints (Req4). LDF does not implement resilient matching techniques such as fuzzy matching (Req2) and does not allow non-authoritative statements about a resource to be found (Req3).

SPARQL allows resources to be found based on text search that (partially) matches literal terms (Req1). More advanced matching and ranking approaches such as string similarity or fuzzy matching are generally not available (Req2). As for the findability of non-authoritative statements, SPARQL has largely the same problems as the other three approaches (Req3). There is also no guarantee that all statements are disseminated by some SPARQL endpoint. Endpoints that are not pointed to explicitly cannot be found by automated means. Empirical studies conclude that many SPARQL endpoints have low availability (Buil-Aranda et al., 2013) (Req4). Figure 24 shows the availability rate for a number of SPARQL endpoints.[7]

FEDERATION is implemented by both LDF and SPARQL, allowing queries to be evaluated over multiple endpoints. Federation is currently unable to implement Web-scale resource search (Req5) since every endpoint has to be included explicitly in the query. This requires the user to enter the Web addresses of endpoints, resulting either in low coverage (Req3) or low serviceability (Req6). Other problems are that the slowest endpoint determines the response time of the entire query and results have to be integrated at the client side. Since Web-wide resource search has to span hundreds of thousands of data sources these problems are not merely theoretical.

Summarizing, findability is a problem on the Semantic Web today. The findability problem will not be solved by implementing existing approaches or standards in a better way, but requires a completely novel approach instead.

7 Source: *Sparqles* (`http://sparqles.ai.wu.ac.at/`). We refer the reader to the Sparqles website for the most recent version of this plot.

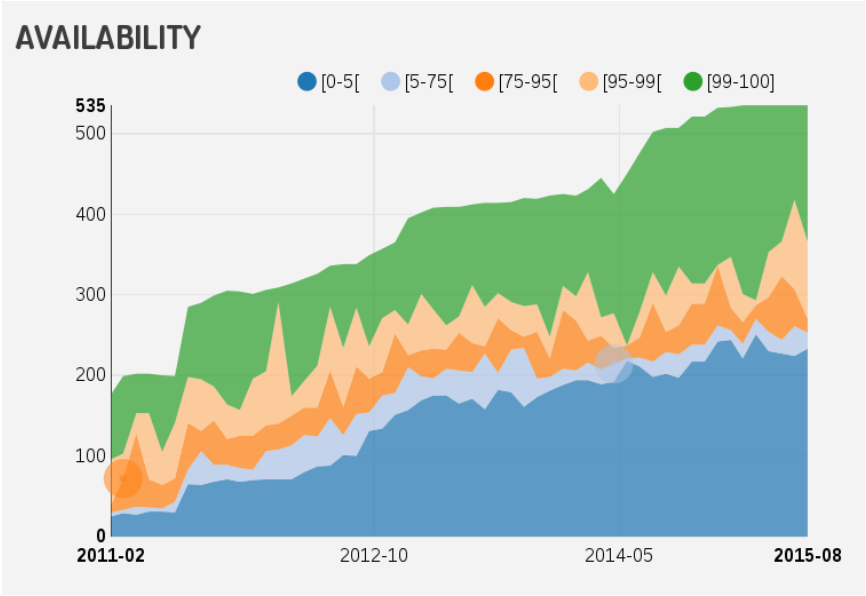


Figure 24: Number of SPARQL endpoints and their availability over time, as measured by Sparqles. The X-axis represents time, whereas the number of endpoints at a given time point is given on the Y-axis. The colors represent availability rates, as indicated in the legend.

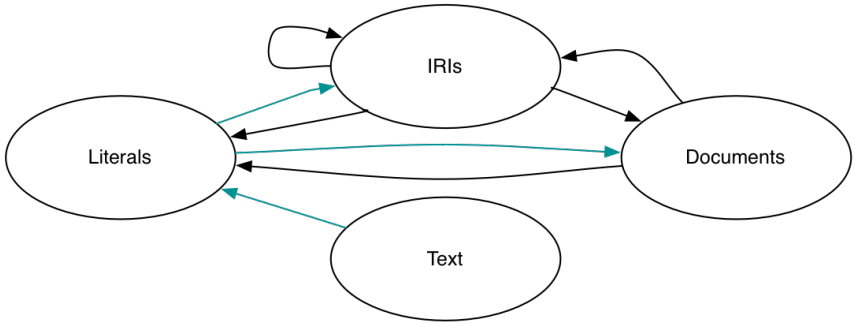


Figure 25: Representation of the capabilities of LOTUS, providing the missing aquamarine link between standard Semantic Web resources and text and Literals

5.3 RELATED WORK

As we discussed in the introduction of this chapter, EL systems do not have a way to simultaneously access a large set of structured knowledge sources through text queries. Notably, this is a common problem in other disambiguation tasks, such as WSD, where it is non-trivial to find resources beyond what is found in

WordNet.[8] While dedicated text indices for a single knowledge base, DBpedia, have been built for the sake of EL systems, like AGDISTIS (Moussallem et al., 2017), textual access to a large set of knowledge bases has not been pursued by the semantic NLP community.

Several systems have implemented text-based search over Semantic Web data: Swoogle (Ding et al., 2004), SemSearch (Lei et al., 2006), Falcons (Cheng et al., 2008), Semplore (Wang et al., 2009), SWSE (Hogan et al., 2011), Hermes (Tran et al., 2009), Sindice/Sigma (Tummarello et al., 2007)[9]. Swoogle allows keyword-based search of Semantic Web documents. Hermes performs keyword-based matching and ranking for schema resources such as classes and (object) properties. Falcons, Semplore, SWSE, and Sindice search for schema and data alike. LOTUS can do much more (see Figure 25): it can relate keywords, IRIs, and documents to each other (in all directions).

Exactly how existing systems extract keywords from RDF data is largely undocumented. Most services use standard NLP approaches such as stemming and stopword removal to improve the keyword index. Furthermore, n-grams are used to take multi-word terms into account (Swoogle), whereas WordNet is used by Hermes to enrich keywords with synonyms. Many systems rely on off-the-shelf text indexing tools such as Lucene that perform similar keyword extraction tasks.

Many systems use metrics derived from the Information Retrieval field such as Term Frequency (TF), Inverse Document Frequency (IDF) and PageRank to rank results. Hermes, for example, implements a TF variant called Element Frequency (EF) that quantifies the popularity of classes/properties in terms of the number of instances they have. Sindice calculates a version of IDF by considering the distinctiveness of data sources from which instances originate, as well as the authority of a data sources: resources that appear in a data source that has the same host name are ranked higher than non-authoritative ones. In practice, TF and IDF are often combined in a single TF/IDF-based metric, in order to have a balanced measure of popularity and distinctiveness.

While the basic idea of adapting existing IR metrics such as TF/IDF and PageRank for text-based search of Semantic Web resources is a common ground for existing systems, they all implement this idea in different ways; using different metrics, the same metrics in different ways, or combining the various metrics in a different way. This makes it difficult to compare existing text-based Semantic Web search systems with one another. Also, the fact that the adapted algorithms differ between these systems provides evidence that there is no single retrieval algorithm that fits every use case. Acknowledging that combining existing approaches into a final ranking over end-results is highly application-dependent and is as much an art as a science, LOTUS takes a very different approach. LOTUS provides an environment in which multiple matching and

8 Resources like BabelNet (Navigli and Ponzetto, 2012) enable this to some extent, by merging WordNet with several other linguistic resources.

9 A dedicated comparison to Sindice can be found in the original publication (Ilievski et al., 2015)

ranking approaches can be developed and combined. This makes it much easier to evaluate the performance of individual rankers on the global end result.

Existing systems operate on data collections of varying size. Sindice, Falcons and Hermes are formally evaluated over hundreds of millions of statements, while Semplore is evaluated over tens of millions of statements. Falcons, Swoogle and Sindice have at some point in time been available as public Web Services for users to query. With Sindice being discontinued in 2014, no text-based Semantic Web search engine is widely available to the Semantic Web community today.

In addition to the work on semantic search engines, there have been multiple attempts to extend existing SPARQL endpoints with more advanced NLP tooling such as fuzzy string matching and ranking over results (Feyznia et al., 2014; Ichinose et al., 2014; Mulay and Kumar, 2011). This improves text search for a restricted number of query endpoints but does not allow text-based queries to cover a large number of endpoints, as the problem of integrating ranked results from many endpoints at the client side has not been solved. Virtuoso's LOD Cache[10] provides public access to a text search-enriched SPARQL endpoint. It differs from LOTUS in that it does not allow the matching and ranking algorithms to be changed or combined in arbitrary ways by the user and as a commercial product its specific internals are not public.

5.4 ACCESS TO ENTITIES AT LOD SCALE WITH LOD LAB

In this section we provide background on the infrastructure and the data collection on top of which we have built our text index, LOTUS.

5.4.1 *LOD Lab*

LOD Laundromat (Beek et al., 2014) is a Semantic Web crawling and cleaning architecture and centralized data collection, available at `http://lodlaundromat.org`. LOD Laundromat spans hundreds of thousands of data documents and tens of billions of ground statements. The collection of datasets that it comprises is continuously being extended.

LOD Lab (Rietveld et al., 2015a) is an infrastructure centered around the LOD Laundromat data collection. The vision of this entire LOD Lab ecosystem is broader, but largely resembles that of LOTUS: to enable the use of the vast amount of structured knowledge found in the LOD Laundromat to generalize evaluations and to enrich applications in various domains. Unfortunately, the focus on one or a handful of knowledge bases is not exclusive to entity linking; most evaluations and applications of Semantic web are based on a small number of well-known knowledge bases. As an illustration, Figure 26 shows that most research papers at the conference ISWC 2014 evaluate on DBpedia.

The obstacles of performing a web-of-data-wide evaluation are numerous and multifaceted, including data collection, quality, accessibility, scalability, availabil-

10 `http://lod.openlinksw.com/`

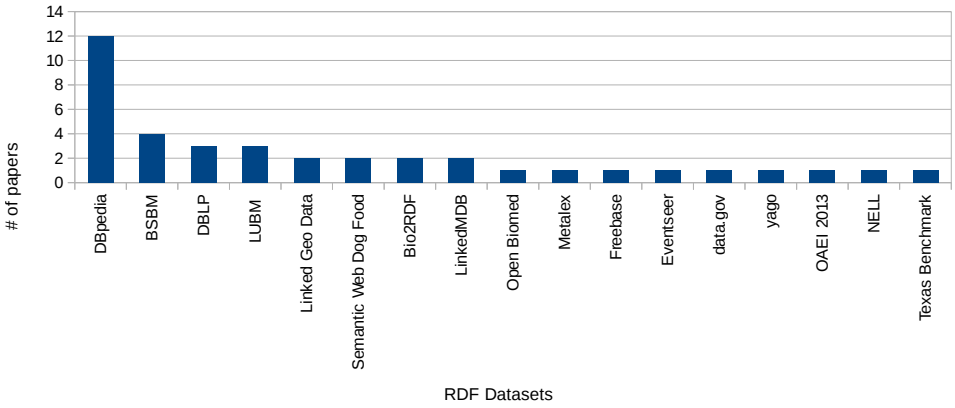


Figure 26: Overview of datasets used in evaluations of papers accepted in the ISWC 2014 research track. For each dataset the number of articles that use it is shown. Originally published in (Rietveld et al., 2015a).

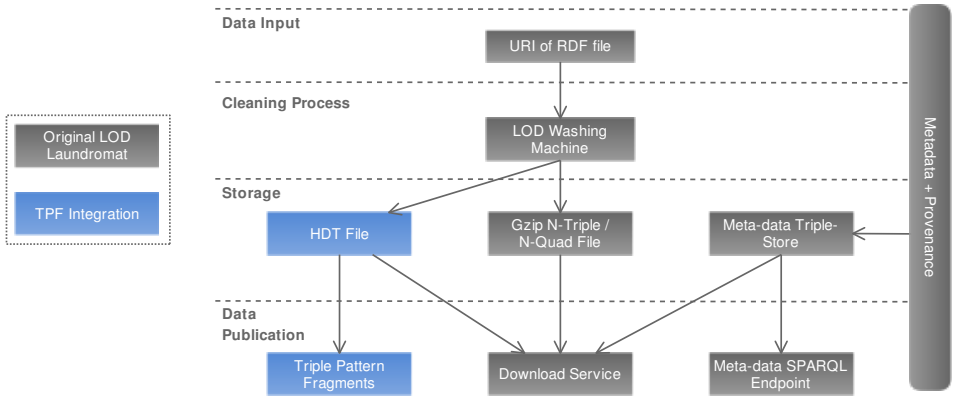


Figure 27: The LOD Laundromat (re)publishing workflow.

ity, and findability. This translates to an almost linear cost of adding one more knowledge source to an evaluation: the knowledge source needs to be collected, cleaned, indexed, etc. Prior to LOD Lab, these steps required human effort. The LOD Laundromat-centered architecture decreases this effort significantly, opening the perspective for large-scale evaluations and applications that take the true variety and richness of LOD into account.

The LOD Laundromat systematically improves the state-of-the art, by systematically crawling, cleaning, indexing, analyzing, and republishing data in a unified way. Next, its family of simple query tools allows researchers and developers to query, access, analyze, and manipulate hundreds of thousands of data documents seamlessly. Each solution is implemented within the LOD Laundromat

framework. Figure 27 shows the LOD Laundromat (re)publishing framework which consists of the following components:[11]

LOD BASKET A list of initial pointers to online available RDF documents, plus the ability for human and machine users to add new pointers.

LOD WASHING MACHINE A full automated cleaning mechanism for RDF documents, implementing a wealth of standards and best practices.

LOD WARDROBE A centralized store of a large subset of the LOD Cloud. Each clean RDF document is made available in several representations and through several Web APIs.

LOD LAUNDROMAT METADATASET Stores metadata about each RDF document's cleaning process, as well as structural metrics about the data graph, together with provenance data about how these metrics are calculated.

INDEX A very large-scale key/value store that maps IRIs and namespaces to documents in which those IRIs and namespaces appear.

LOTUS A full-text search index on all textual RDF literals that allows resources, statements about resources, to be found through a configurable combination of matchers and rankers. We will describe LOTUS in detail in this chapter.

LOD LAB An approach to reproducing Linked Data research which uses the various components in an integrated way.

5.4.2 *APIs and tools*

While the LOD Lab base infrastructure is quite complex, using the LOD Lab to scale Semantic Web evaluation is much simpler. LOD Lab relies on the following Web APIs the LOD Laundromat makes available:

- **LOTUS** Search for IRIs and triples based on free text matching and filtering.
- **Linked Data Fragments** (Verborgh et al., 2014) (LDF) Search for triples based on (incomplete) patterns.
- **Namespace index** Search for documents wherein a given namespace occurs.
- **IRI index** Search for documents in which a given IRI occurs.
- **SPARQL** Search for metadata about the crawling process and (structural) properties of the data to filter documents.

11 This thesis section is based on (Beek et al., 2017). We refer the reader to this paper for more detailed description of the underlying problems and the LOD Lab method of addressing them.

- **Datadump** Bulk access to full RDF documents.

In the case of LOTUS, before we query the entire LOD Laundromat data collection someone needs to first request the list of all documents from the SPARQL endpoint and then pose an LDF request for each document. Use cases like this are handled by *Frank* (Beek and Rietveld, 2015), a Command-Line Interface (CLI) that allows the most common operations to be performed more easily. *Frank* is implemented as a single-file Bash script. As with other Unix-inspired tools, *Frank* can be easily combined with other commands through Bash pipes. This makes it relatively easy to run large-scale evaluations over the LOD Laundromat data collection. In fact, the three evaluations that were (re)run in the original LOD Lab paper (Rietveld et al., 2015b) all consisted of only a couple of calls to *Frank*.

These are the four main LOD Lab tasks that can be performed by using the LOD Laundromat set of tools and APIs:

1. Find IRIs with LOTUS.
2. Find statements with LDF or Frank.
3. Find documents, by: querying the SPARQL endpoint, the document index directly, or with Frank. Metadata on documents can be obtained with the *frank meta* command.
4. Combine finding IRIs, statements and documents with other tools, and custom algorithms.

LOD Lab provides added value for a wide range of use cases, such as data search, publishing support, and web-wide evaluation of algorithms.

5.5 LOTUS

LOTUS relates unstructured to structured data using RDF as a paradigm to express such structured data. It indexes natural text literals that appear in the object position of RDF statements in LOD Laundromat (see section 5.4 for a detailed description of LOD Laundromat and LOD Lab) and allows the denoted resources to be findable based on approximate matching. LOTUS currently includes four different matching algorithms and eight ranking algorithms, which leverage both textual features and relational information from the RDF graph.

5.5.1 *Model*

Denoted resources RDF defines a graph-based data model in which resources can be described in terms of their relations to other resources. An RDF statement expresses that a certain relation holds between a pair of resources.

The textual labels denoting some of these resources provide an opening to relate unstructured to structured data. LOTUS does not allow every resource in the Semantic Web to be found through text search, as some resources are not

denoted by a term that appears as a subject of a triple whose object term is a textual label. Fortunately, many Semantic Web resources are denoted by at least one textual label and as the Semantic Web adheres to the Open World Assumption, resources with no textual description today may receive one tomorrow, as everyone is free to add new content.

RDF Literals In the context of RDF, textual labels appear as part of *RDF literals*. An RDF literal is either a pair $\langle \mathrm{D}, \mathrm{LEX} \rangle$ or a triple $\langle \texttt{rdf:langString}, \mathrm{LEX}, \mathrm{LT} \rangle$ (Cyganiak et al., 2014). D is a datatype IRI denoting a datatype. LEX is a Normal Form C (NFC) Unicode string (Davis and Whistler, 2012). LT is a language tag identifying an IANA-registered natural language per BCP 47 (Phillips and Davis, 2009). Semantically speaking, RDF literals denote resources, similar to the way in which IRIs denote resources. A datatype D defines the collection of allowed lexical forms (**lexical space**), the collection of resources denoted by those lexical forms (**value space**), a functional mapping from the former to the latter and a non-functional mapping from the latter to the former.

We are specifically interested in literals that contain natural language text. However, not all RDF literals express – or are intended to express – natural language text. For instance, there are datatype IRIs that describe a value space of date-time points or polygons. Even though each dataset can define its own datatypes, we observe that the vast majority of RDF literals use RDF or XSD datatypes. This allows us to circumvent the theoretical limitation of not being able to enumerate all textual datatypes and focus on the datatypes `xsd:string` and `rdf:langString` (Cyganiak et al., 2014). Unfortunately, in practice we find that integers and dates are also regularly stored under these datatypes. As a simple heuristic filter LOTUS only considers literals with datatype `xsd:string` and `xsd:langString` that contain at least two consecutive alphabetic Unicode characters.

5.5.2 *Language tags*

The language tag assigned by the original data publishers can consist of multiple, concatenated subtags. Since our language detection tools only provide an ISO 639-2 two-character language code in most cases, we focus our comparison on the primary language subtag, i.e., the first two characters of each language-tagged string. Another motivation for enforcing this abstraction is that it is more difficult to distinguish fine-grained language differences from a semantic point of view. For instance, if the original data publisher supplied language tag `de-DE`, it is difficult to determine whether `de-AU` or de would also have been correct annotations. The granularity level that we choose, two-character primary language tags, is satisfactory for identifying most languages, although there are exceptional cases in which the secondary language subtag is also required for denoting the language. Most notably, this is the case for Chinese languages where `zh-CN` denotes a language that is different from `zh-TW`.

Table 30: The distribution of language tags in the LOD Laundromat data collection

Language tag	Occurrences
en	878,132,881
de	145,868,558
fr	129,738,855
it	104,115,063
es	82,492,537
ru	77,856,452
nl	75,226,900
pl	59,537,848
pt	56,426,484
sv	47,903,859
other language tag	607,012,252
XSD string	1,281,785,207
textual literals	3,544,028,391

We analyzed the language tags of the literals found in the LOD Laundromat, identified by the heuristic described in the Section 5.5.1.[12] The obtained results are shown in Table 30. We observe that 63.83% (2.26 billion) of all literals contain an explicit language tag assigned by the data publisher. Within this set, the most language-tagged literals are in English, followed by German, French, Italian and Spanish. This shows that Linked Data contains a strong representation bias towards languages of European origin, with the 10 most frequent language tags representing European languages. 73.26% of all language-tagged literals belong to one of the 10 most frequently occurring languages.

Comparing three libraries for automatic language detection, we found that the Compact Language Detection (CLD)[13] library performs best consistently. This library can reportedly distinguish between 160 languages. In the current version of LOTUS we perform an automatic language tagging using CLD for the remaining 36.17% literals that have no manually assigned language tag. Table 31 shows the top-10 most-frequently assigned tags on the literals without an original language tag.

In some instances a lexical form can be correctly annotated with multiple primary language tags. This is especially true for proper names – these often share the same surface form in a plurality of languages. For instance, what is the right language tag for "Amsterdam": English, Dutch, or a set of languages? Ideally, the world languages would be grouped into a hierarchy of language tags, thus allowing the data publisher to specify a group or category of similar language tags (e.g., all Germanic languages). This representation is not standardized at the moment.

12 We refer the reader to the original paper for further details (Beek et al., 2018).

13 `https://github.com/dachev/node-cld`

Table 31: Languages among the untagged strings according to CLD

Language tag	Occurrences
en	348,262,051
nl	27,700,617
de	25,166,990
da	12,574,645
ja	9,158,424
es	8,593,138
fr	7,248,383
nn	7,114,156
el	4,323,837
la	2,873,684

5.5.3 *Linguistic entry point to the LOD Cloud*

Inherent to its integration with the LOD Laundromat architecture, LOTUS fulfills three of the requirements we set in Section 5.2: 1. the scale on which LOTUS operates is in range of billions and is 100 times bigger than that of previous semantic search systems that were made generally available for the community to use (Req5); 2. since the LOD Laundromat data is collected centrally, finding both authoritative and non-authoritative RDF statements is straightforward (Req3); 3. as a cached copy of linked data, LOD Laundromat allows its IRIs to be dereferenceable even when the original sources are unavailable (Req4).

LOTUS allows RDF statements from the LOD Laundromat collection to be findable through approximate string matching on natural language literals. Approximate string matching (Navarro, 2001) is an alternative to exact string matching, where one textual pattern is matched to another while still allowing a number of errors. In LOTUS, query text is approximately matched to existing RDF literals (and their associated documents and IRI resources) (Req1).

In order to support the approximate matching and linguistic access to LD through literals, LOTUS makes use of an inverted index. As indexing of big data in the range of billions of RDF statements is expensive, the inverted index of LOTUS is created offline. This also allows the approximation model to be efficiently enriched with various precomputed retrieval metrics.

5.5.4 *Retrieval*

Matching algorithms Approximate matching of a query to literals can be performed on various levels: as phrases, as sets of tokens, or as sets of characters. LOTUS implements four matching functions to cope with this diversity (Req2):

M1. PHRASE MATCHING: Match a phrase in an object string. Terms in each result should occur consecutively and in the same order as in the query.

M2. DISJUNCTIVE TOKEN MATCHING: Match some of the query tokens in a literal. The query tokens are connected by a logical "OR" operator, expressing that each match should contain at least one of the queried tokens. The order of the tokens between the query and the matched literals need not coincide.

M3. CONJUNCTIVE TOKEN MATCHING: Match all query tokens in a literal. The set of tokens are connected by a logical "AND" operator, which entails that all tokens from the query must be found in a matched literal. The order of the tokens between the query and the matched literals need not coincide.

M4. CONJUNCTIVE TOKEN MATCHING WITH EDIT DISTANCE: Conjunctive matching (with logical operator "AND") of a set of tokens, where a small Levenshtein-based edit distance on a character level is permitted. This matching algorithm is intended to account for typos and spelling mistakes.

To bring a user even closer to her optimal set of matches, LOTUS facilitates complementary filtering based on the language of the literals, as explicitly specified by the dataset author or automatically detected by a language detection library. While a language tag can contain secondary tags, e.g. to express country codes, LOTUS focuses on the primary language tags which denote the language of a literal and abstracts from the complementary tags.

Ranking algorithms Ranking algorithms on the Web of data operate on top of a similarity function, which can be content-based or relational (Christophides et al., 2015).[14] The content-based similarity functions exclusively compare the textual content of a query to each potential result. Such comparison can be done on different granularity of text, leading to character-based (Levenshtein similarity, Jaro similarity, etc.) and token-based (Jaccard, Dice, Overlap, Cosine similarity, etc.) approaches. The content similarity function can also be information-theoretical, exploiting the probability distributions extracted from data statistics. Relational similarity functions complement the content similarity approaches by considering the underlying structure of the tree (tree-based) or the graph (graph-based).

We use this classification of similarity algorithms as a starting point for our implementation of three content-based (R1-R3) and five relational functions (R4-R8) in LOTUS, thus addressing Req7:

R1. CHARACTER LENGTH NORMALIZATION: The score of a match is counter-proportional to the number of characters in the lexical form of its literal.

R2. PRACTICAL SCORING FUNCTION (PSF): [15] The score of a match is a product of three token-based information retrieval metrics: term frequency

14 The reader is referred to this book for detailed explanation of similarity functions and references to original publications

15 This function is the default scoring function in ElasticSearch. Detailed description of its theoretical basis and implementation is available at `https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html`

(TF), inverse-document frequency (IDF) and length normalization (inverse-proportional to the number of tokens).

R3. PHRASE PROXIMITY: The score of a match is inverse-proportional to its edit distance with respect to the query.

R4. TERMINOLOGICAL RICHNESS: The score of a match is proportional to the presence of controlled vocabularies, i.e. classes and properties, in the original document from which the RDF statement stems from.

R5. SEMANTIC RICHNESS OF THE DOCUMENT: The score of a match is proportional to the mean graph connectedness degree of the original document.

R6. RECENCY RANKING: The score of a match is proportional to the moment in time when the original document was last modified. Statements from recently updated documents have higher score.

R7. DEGREE POPULARITY: The score of a match is proportional to the total graph connectedness degree (indegree + outdegree) of its subject resource.

R8. APPEARANCE POPULARITY: The score of a match is proportional to the number of documents in which its subject appears.

In the next section, the implementation of LOTUS is detailed.

5.6 IMPLEMENTATION

The LOTUS system architecture consists of two main components: the Index Builder (IB) and the Public Interface (PI). The role of the IB is to index strings from LOD Laundromat; the role of the PI is to expose the indexed data to users for querying. The two components are executed sequentially: data is indexed offline, after which it can be queried via the exposed public interface.

5.6.1 *System architecture*

Since our rankings rely on metadata about documents and resources which is reused across statements, we need clever ways to compute and access this metadata. For this purpose, we pre-store the document metadata needed for the ranking algorithms R4-R6, which includes the last modification date, the mean graph degree and the terminological richness coefficient of each LOD Laundromat document. To obtain these, we use the LOD Laundromat access tools that were described in Section 5.4: Frank (Beek and Rietveld, 2015) and the SPARQL endpoint[16]. The rankings R7 and R8 use metadata about resource IRIs. Computing, storage, and access to this information is more challenging, as the number of resources in the LOD Laundromat is huge and their occurrences are scattered

16 http://lodlaundromat.org/sparql/

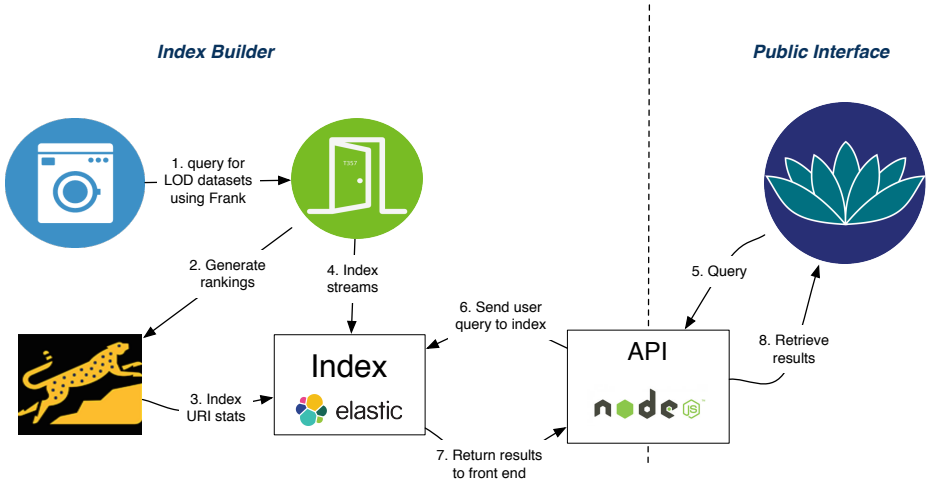


Figure 28: LOTUS system architecture

across documents. To resolve this, we store the graph degree of a resource and number of documents where it appears in RocksDB.[17]

Once the relational ranking data is cached, we start the indexing process over all data from LOD Laundromat through a batch loading procedure. This procedure uses LOD Laundromat's query interface, Frank (Step 1 in Figure 28), to list all LOD Laundromat documents and stream them to a client script. Following the approach described in Section 5.5, we consider only the statements that contain a natural language literal as an object. The client script parses the received RDF statements and performs a bulk indexing request in ElasticSearch (ES),[18] where the textual index is built (Steps 2, 3 and 4 in Figure 28).

As soon as the indexing process is finished, LOTUS contains the data it needs to perform text-based retrieval over the LOD Laundromat collection. Its index is only incrementally updated when new data is added in LOD Laundromat.[19]

For each RDF statement from the LOD Laundromat, we index: 1) *Information from the statement itself:* subject IRI, predicate IRI, lexical form of the literal ("string"), length of the "string" field (in number of characters), language tag of the literal and document ID; 2) *Metadata about the source document:* last modification date, terminological richness coefficient and semantic richness coefficient; 3) *Metadata about the subject resource:* graph degree and number of documents in which the resource appears.

We store the metadata for 2) and 3) in a numeric format to enable their straightforward usage as ranking scores by ElasticSearch. Each ElasticSearch entry has the following format:

17 `http://rocksdb.org/`
18 `https://www.elastic.co/products/elasticsearch`
19 This procedure is triggered by an event handler in the LOD Laundromat itself.

```
{
  "docid": IRI,
  "langtag": STRING,
  "predicate": IRI,
  "string": STRING,
  "subject": IRI,
  "length": float,
  "docLastModified": float,
  "docTermRichness": float,
  "docSemRichness": float,
  "uriDegree": int,
  "uriNumDocs": int
}
```

In the current version of LOTUS, the fields "subject", "predicate", and "string" are all *analyzed* (preprocessed by ElasticSearch), which allows them to be matched in a flexible manner. Subjects and predicates can therefore be matched based on substring queries - for instance, the user can query for every subject that contains `sw.open.cyc`, or every predicate that does not contain the word `label`.

The field "string" can be matched on some or all of its tokens. For instance, querying for "California" returns both `http://dbpedia.org/page/California` through an exact match and `http://data.semanticweb.org/organization/information-sciences-institute-university-of-southern-california` through a partial match to its label "ISI / University of Southern California".

We next discuss the implementation of the different options for matching and ranking.

5.6.2 *Implementation of the matching and ranking algorithms*

While the matching algorithms we introduce are mainly adaptations of off-the-shelf ElasticSearch string matching functions, we allow them to be combined with relational information found in the RDF statement to improve the effectiveness of the matching process. The approximate matching algorithms M1-M4 operate on the content of the "string" field, storing the lexical form of a literal. This field is preprocessed ("analyzed") by ElasticSearch at index time, thus allowing existing ElasticSearch string matching functionalities to be put into practice for matching. We allow users to further restrict the matching process by specifying relational criteria: language tag of the literal ("langtag"), associated subject and predicate as well as LOD Laundromat document identifier.

Similarly to the matching algorithms, our ranking algorithms rely on both ElasticSearch functionality and relational information extracted from LOD Laundromat. Concretely, our rankings are based on: 1) *Scoring functions from ElasticSearch* (R1-R3); 2) *Document-level scores:* last modification date (R4), terminological rich-

ness (R5) and semantic richness (R6); 3) *Resource-level scores:* graph degree (R7) and number of documents that contain the resource (R8).

5.6.3 *Distributed architecture*

In our implementation, we leverage the distributed features of ElasticSearch and scale LOTUS horizontally over 5 servers. Each server has 128 GB of RAM, 6 core CPU with 2.40GHz and 3 SSD hard disks with 440 GB of storage each. We enable data replication to ensure high runtime availability of the system.

The index of LOTUS contains 4.33 billion entries, allowing over two billion distinct LOD Laundromat URIs to be found. Because the indexes of LOTUS are very large, they have to be created with Big Data tools, specifically Hadoop[20] and RocksDB.[21] Cached versions of the data are stored in Header Dictionary Triples (HDT)[22] and are exposed through a Linked Data Fragments (LDF)[23] API. Metadata about the source documents is stored in a Virtuoso triple store and exposed through a SPARQL endpoint. Because all components of the LOTUS framework are exposed using standards-compliant web APIs, it is easy for developers to extend the functionality of LOTUS.

5.6.4 *API*

Users can access the underlying data through an API. The usual query flow is described in steps 5-8 of Figure 28. We expose a single query endpoint,[24] through which the user can supply a query, choose a combination of matching and ranking algorithms, and optionally provide additional requirements, such as language tag or number of results to retrieve. The basic query parameters are:[25]

STRING A natural language string to match in LOTUS

MATCH Choice of a matching algorithm, one of *phrase (M1), terms (M2), conjunct (M3), fuzzyconjunct (M4)*

RANK Choice of a ranking algorithm, one of *lengthnorm (R1), psf (R2), proximity (R3), termrichness (R4), semrichness (R5), recency (R6), degree (R7), appearance (R8)*

SIZE Number of best scoring results to be included in the response

LANGTAG Two-letter language identifier

20 http://hadoop.apache.org
21 http://rocksdb.org
22 http://www.rdfhdt.org/
23 http://linkeddatafragments.org/
24 http://lotus.lodlaundromat.org/retrieval
25 See http://lotus.lodlaundromat.org/docs for additional parameters and more detailed information.

LOTUS is also available as a web interface at `http://lotus.lodlaundromat.org/` for human-friendly exploration of the data, thus fulfilling Req6 on usefulness for both humans and machines. Code of the API functions and data from our experiments can be found on GitHub.[26] The code used to create the LOTUS index is also publicly available.[27]

5.6.5 *Examples*

The LOTUS Web UI can be found at `http://lotus.lodlaundromat.org`. A query is issued by filling in the respective HTML forms. Different matchers and rankers can be chosen from dropdown menus. The same queries that are issued through the Web UI can also be performed algorithmically. The URI-encoded query is shown inside the Web UI. For instance, the following searches for RDF terms and statements about monkeys:

```
http://lotus.lodlaundromat.org/retrieve?string=monkey
```

For each option that is changed in the Web UI, the URI changes as well. For instance, the following excludes results where the subject term is not an IRI (i.e., excluding blank nodes).

```
http://lotus.lodlaundromat.org/retrieve?string=monkey&\
noblank=true
```

LOTUS is not limited to searching textual RDF literals, descriptive text often appears in IRIs as well. The following query searches for monkeys that are defined in OpenCyc:

```
http://lotus.lodlaundromat.org/retrieve?string=monkey&\
noblank=true&subject=sw.opencyc.org
```

In the following query we exclude results where 'label' (e.g., `rdfs:label`) appears in the predicate position:

```
http://lotus.lodlaundromat.org/retrieve?string=monkey&\
noblank=true&subject=sw.opencyc.org&predicate=NOT%20label
```

Another useful feature is to filter for strings whose content belongs to a particular natural language. The following only returns language-tagged strings that belong to the English language (language tag `en`):

```
http://lotus.lodlaundromat.org/retrieve?string=monkey&\
noblank=true&subject=sw.opencyc.org&predicate=NOT%20label&\
langtag=en
```

These sample queries in LOTUS show that the various filters can be easily combined to make fine-grained queries.

26 https://github.com/filipdbrsk/LOTUS_Search/
27 https://github.com/filipdbrsk/LOTUS_Indexer/

5.7 PERFORMANCE STATISTICS AND FLEXIBILITY OF RETRIEVAL

As LOTUS does not provide a one-size-fits-all solution, we present some performance statistics and scenarios in this section. We test LOTUS on a series of queries and show the impact of different matching and ranking algorithms. Both scalability and flexibility are a crucial component of a text index that aspires to enable linking of long-tail entities. By linking to entities on a much larger scale, we automatically increase the size of the seed set of entities considered. Knowing that most entities are tail entities, this increase in size and computational complexity could in theory be very large.[28] Flexible retrieval is important because the matching between surface forms and instances tends to be case-dependent - finding forms that contain a spelling error would require a different retrieval mechanism compared to forms that should be matched strictly as a full phrase, or forms that are to be treated as a set of tokens.

Table 32: Statistics on the indexed data of LOTUS.

total # literals encountered	12,380,443,617
#xsd:string literals	6,205,754,116
#xsd:langString literals	2,608,809,608
# indexed entries in ES	4,334,672,073
# distinct sources in ES	493,181
# hours to create the ES index	67
disk space used for the ES index	509.81 GB
# in_degree entries in RocksDB	1,875,886,294
# out_degree entries in RocksDB	3,136,272,749
disk space used for the RocksDB index	46.09 MB

5.7.1 *Performance statistics*

Statistics over the indexed data are given in Table 32. LOD Laundromat contains over 12 billion literals, 8.81 billion of which are defined as a natural language string (`xsd:string` or `xsd:langString` datatype). According to our approach, 4.33 billion of these (~35%) express natural language strings, stemming from 493,181 distinct datasets.[29]. The LOTUS index was created in 67 hours, consuming 509.81 GB of disk space.

To indicate the performance of LOTUS from a client perspective we performed 324,000 text queries. We extracted the 6,000 most frequent bigrams, trigrams, and quadgrams (18,000 N-grams in total) from the source of *A Semantic Web Primer* (Antoniou et al., 2012). Non-alphabetic characters were first removed. For each

28 In practice, most tail entities are still not found in the LOD cloud, but that could change.

29 The number of distinct sources in LOTUS is lower than the number of documents in LOD Laundromat, as not every document contains natural language literals.

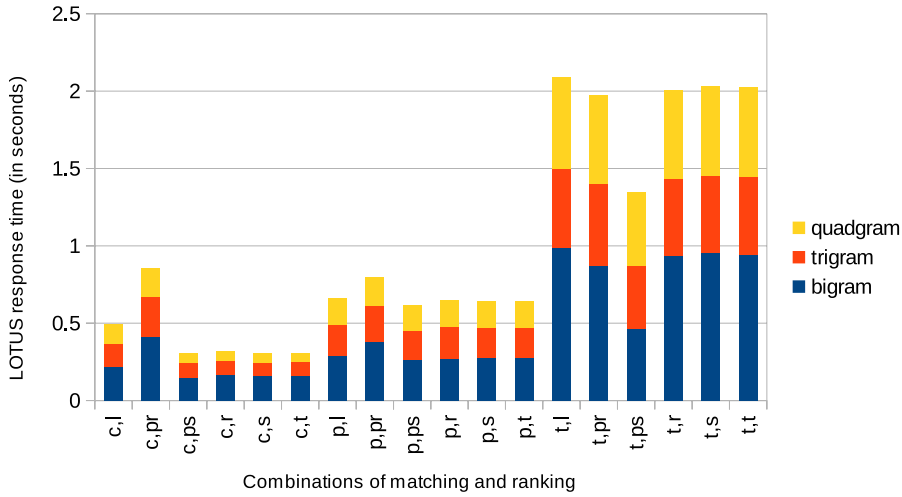


Figure 29: LOTUS average response times in seconds for bi-, tri-, and quadgram requests. The horizontal axis shows 18 combinations of a matcher and a ranker. The matchers are *conjunct* (c), *phrase* (p), and *terms* (t). The rankers are *length normalization* (l), *proximity* (pr), *psf* (ps), *recency* (r), *semantic richness* (s), and *term richness* (t). The bar chart is cumulative per match+rank combination. For instance, the first bar indicates that the combination of conjunct matching and length normalization takes 0.20 seconds for bigrams, 0.15 seconds for trigrams, 0.15 seconds for quadgrams and 0.5 seconds for all three combined. The slowest query is for bigrams with terms matching and length normalization, which takes 1.0 seconds on average.

N-gram we performed a text query using one of three matchers in combination with one of six rankers. The results are shown in Figure 29. We observe certain patterns in this Figure. Matching disjunctive terms (M2) is strictly more expensive than the other two matching algorithms. We also notice that bigrams are more costly to retrieve than trigrams and quadgrams. Finally, we observe that there is no difference between the response time of the relational rankings which is expected, because these rank results in the same manner, through sorting prestored integers in a decreasing order.

The performance and scalability of LOTUS is largely due to the use of ElasticSearch, which justifies our rationale in choosing ElasticSearch as one of our two main building blocks: it allows billions of entries to be queried within reasonable time constraints, even running on an academic hardware infrastructure.

5.7.2 *Flexibility of retrieval*

To demonstrate the flexibility and the potential of the LOTUS framework, we performed retrieval on the query "graph pattern". We matched this query as a phrase (M1) and iterated through the different ranking algorithms. The top results obtained with two of the different ranking modes are presented in Figure 30.

The different ranking algorithms allow a user to customize her results. For example, if a user is interested in analyzing the latest changes in a dataset, the Recency ranking algorithm will retrieve statements from the most recently updated datasets first. A user who is more interested in linguistic features of a query can use the length normalization ranking to explore resources that match the query as precisely as possible. Users interested in multiple occurrences of informative phrases could benefit from the practical scoring function. When the popularity of resources is important, the degree-based rankings can be useful.

Users can also vary the matching dimension. Suppose one is interested to explore resources with typos or spelling variation: fuzzy conjunctive matching would be the appropriate matching algorithm to apply.

5.8 FINDING ENTITIES BEYOND DBPEDIA

The research question that is being addressed in this chapter is *How to improve the access to tail entities beyond DBpedia?* We have built LOTUS to enable access to a much wider body of knowledge than the general-domain entities in DBpedia. We thus hope that the results of LOTUS include contextually-relevant, tail entities, as found in the underlying LOD cloud.

In this section, we illustrate the potential of LOTUS to enable LOD-wide textual access to knowledge beyond DBpedia through three Entity Linking scenarios. We perform a small recall evaluation on a standard benchmark dataset, namely the AIDA-YAGO2 entity linking benchmark dataset (Hoffart et al., 2011). We also present two domain-specific use cases, namely Local monuments and Scientific journals, which we expect to contain a larger number of NILs and tail entities.

For each use case scenario, we gather a set of entities and query each entity against LOTUS. We used the following matching options: M1 (phrase matching), M1 + LT (M1 with a language tag), M2 (disjunctive term matching), and M2 + LT (M2 with a language tag). We ranked the results using the default ElasticSearch ranker, PSF. We counted the amount of entities without results and the proportion of DBpedia resources in the first 100 candidates, as a comparison to the (currently) most popular knowledge base. We then inspected a number of query results to assess their relevance to the search query. In the remainder of this section we detail the specifics of each use case.[30]

30 The experiments presented here were executed with an earlier version of the LOTUS index, presented in (Ilievski et al., 2015). Repeating this experiment with the current LOTUS index might result in slightly different numbers.

Table 33: Use case statistics on AIDA-YAGO2, showing: # of queries, # queries for which no result is retrieved, # queries for which we only find resources other than DBpedia, and proportion of DBpedia resources in the first 100 results per query type. Matching options: M1 (phrase matching), M1 + LT (M1 with a language tag `en`), M2 (disjunctive term matching), and M2 + LT (M2 with a language tag `en`). Ranker: R3 (Practical Scoring Function).

	M1	M1 + LT	M2	M2 + LT
# queries	5,628			
# no result	1,286	1,286	54	54
# no DBpedia	1,723	1,701	185	115
% DBpedia	69.49	77.68	64.35	79.11

5.8.1 *AIDA-YAGO2*

The AIDA-YAGO2 dataset (Hoffart et al., 2011) is an extension of the CoNLL 2003 Named Entity Recognition Dataset (Tjong Kim Sang and Meulder, 2003) to also include links to Wikipedia entries for each entity. 7,112 of the entity phrases in AIDA-YAGO2 have no DBpedia entry. We removed the duplicate entities in each article, providing us with 5,628 entity mentions. We focus on these to show the impact of having access to multiple datasets.

The results of our analysis on AIDA-YAGO2 are given in Table 33. We suspect that the growth in DBpedia since the release of this dataset has improved recall on the named entities, but there is still a benefit of using multiple data sources. We observe that between 64.35% (for matcher M2) and 79.11% (for M2 + LT) of all results stem from DBpedia. For both M1 and M2, the percentage of DBpedia results increases when we include a language tag in the query: this signals that, on average, DBpedia includes language tags for its literals more often than a random non-DBpedia IRI in LOTUS.

The results show that it is useful to look for entities beyond DBpedia, especially when DBpedia gives no result. Query M2 provides results for 71% of the queries that have no DBpedia result. This is the highest increase of recall among the four matching options; the lowest (24%) is measured for M2 + LT.

We also manually inspected the top results of individual queries. For smaller locations, such as the “Chapman Golf Club", relevant results are found in for example `http://linkedgeodata.org/About`. Also, the fact that the different language versions of DBpedia are plugged in helps in retrieving results from localized DBpedias such as for “Ismaïl Boulahya", a Tunisian politician described in `http://fr.dbpedia.org/resource/Isma%C3%AFl_Boulahya`. Some of the retrieval is hampered by newspaper typos, such as “Allan Mullally” (“Alan Mullally” is the intended surface form).

Table 34: Use case statistics on Local Monuments, showing: # of queries, # queries for which no result is retrieved, # queries for which we only find resources other than DBpedia, and proportion of DBpedia resources in the first 100 results per query type. Matching options: M1 (phrase matching), M1 + LT (M1 with a language tag `en`), M2 (disjunctive term matching), and M2 + LT (M2 with a language tag `en`). Ranker: R3 (Practical Scoring Function).

	M1	M1 + LT	M2	M2 + LT
# queries	191			
# no result	23	23	3	3
# no DBpedia	26	25	4	3
% DBpedia	70.48	83.23	67.19	84.92

5.8.2 *Local monuments guided walks*

The interest in applications such as Historypin (`http://www.historypin.org`) or the Rijksmuseum API (`https://www.rijksmuseum.nl/en/api`) shows that there are interesting use cases in cultural heritage and local data. To explore the coverage of this domain in the LOD Laundromat, we created the local monuments dataset by downloading a set of guided walks from the Dutch website `http://www.wandelnet.nl`. We specifically focused on the tours created in collaboration with the Dutch National Railways as these often take walkers through cities and along historic and monumental sites. From the walks 'Amsterdam Westerborkpad' and 'Mastbos Breda', a human annotator identified 112 and 79 entities respectively. These are mostly monuments such as 'De Grote Kerk' (*The big church*) or street names such as 'Beukenweg' (*Beech street*).

Table 34 shows that most queries have results in DBpedia, and the handful of queries that have no result in DBpedia also do not have a result overall in the LOD Laundromat. Namely, considering M2 only 4 out of 191 queries do not result in at least one DBpedia IRI, and only 1 of these 4 has a result elsewhere in the LOD Laundromat. The trend is similar for M1 and M1 + LT: for these matchers, there are overall more cases without a DBpedia result but these queries typically (23 out of 26 times) also have no result elsewhere on the LOD cloud. The percentage of DBpedia results in the top-100 is relatively high for the 'Local Monuments' dataset, ranging between 67.19% for M2 and 84.92% for M2 + LT. Again, we observe that the percentage of DBpedia IRIs increases when we include a language tag in the query to LOTUS.

We manually inspected the top-10 results on a number of queries. Here we find that the majority of the highest ranking results is still coming from DBpedia. However, when no DBpedia link is available, often a resource from the Amsterdam Museum (`http://semanticweb.cs.vu.nl/lod/am/`) or Wikidata (`http://www.wikidata.org`) is retrieved. The focus on entertainment in DBpedia is also shown here for the query 'Jan Dokter', the person who first walked the route to commemorate his family that died in WWII. 'Jan' is a very common

Table 35: Use case statistics on Journals, showing: # of queries, # queries for which no result is retrieved, # queries for which we only find resources other than DBpedia, and proportion of DBpedia resources in the first 100 results per query type. Matching options: M1 (phrase matching), M1 + LT (M1 with a language tag), M2 (disjunctive term matching), and M2 + LT (M2 with a language tag). Ranker: R3 (Practical Scoring Function). Language tag queries were not ran because the dataset is multilingual.

	M1	M1 + LT	M2	M2 + LT
# queries	231			
# no result	10	X	0	X
# no DBpedia	49	X	15	X
% DBpedia	24.83	X	22.33	X

Dutch first name, and 'Dokter' means 'doctor', which results in many results about characters in Dutch and Flemish soap operas who happen to be doctors. This expresses a need for allowing more context to be brought into the search query to filter results better.

Summarizing, the potential of LOTUS to increase the recall on the Local Monuments dataset is notably smaller than on AIDA-YAGO2.

5.8.3 *Scientific journals*

Whitelists (and blacklists) of scientific journals are used by many institutions to gauge the output of their researchers. They are also used by researchers interested in the scientific social networks. One such list is made publicly available by the Norwegian Social Science Data Services Website (`http://www.nsd.uib.no/`). Their level 2 publishing channel contains 231 titles of journals. The majority of these titles is in English, but it also contains some German and Swedish titles barring the use of the language tag in querying.

Table 35 shows the obtained results. As the queries are generally longer and contain more context-specific terms such as "journal", "transactions", "methods", and "association", the query results are generally more relevant. The exception here are the more generic titles, such as "Transportation": these yield, as expected, more generic results.

We observed that only 22-25% of all query results come from DBpedia. In addition, 39 out of 49 queries that have no DBpedia result for M1, and all 15 queries for M2, have a result in LOTUS found in a non-DBpedia knowledge source. Such sources are: ZDB (`http://dispatch.opac.dnb.de/LNG=DU/DB=1.1/`), the 2001 UK's Research Assessment Exercise as exposed through RKB Explorer (`http://rae2001.rkbexplorer.com/`), Lobid (`http://lobid.org/`), and Wikidata.

We can conclude that long-tail entities in scientific journals are especially well-suited for entity linking over the LOD cloud, as they are often represented in more specialized knowledge bases, such as ZDB or RKB Explorer.

5.9 DISCUSSION AND CONCLUSIONS

The lack of a global LOD-wide text index is prohibitive for building LOD-wide entity linkers. Recognizing this, we presented LOTUS, a full-text entry point to the centralized LOD Laundromat collection. We detailed the specific difficulties in accessing textual content in the LOD cloud today and the approach taken by LOTUS to address these. LOTUS allows its users to customize their own retrieval method by exposing analytically well-understood matching and ranking algorithms, taking into account both textual similarity and certain structural properties of the underlying data. LOTUS currently provides 32 retrieval options to be used in different use cases.

We demonstrated the potential of LOTUS to find long-tail entities beyond DBpedia in three small use case scenarios. We were able to find entity representations for the majority of the queries relevant for these use cases. Compared to using only DBpedia, we noted the highest increase of recall by LOTUS on the 'Journals' dataset. Namely, less than a quarter of all results for journals stemmed from DBpedia, and all of the queries without a DBpedia candidate, had a result elsewhere in the LOD Laundromat. We observed a moderate increase of recall on AIDA-YAGO2: up to 36% of the results stemmed from outside DBpedia, and up to 71% of the DBpedia NILs (queries without a result in DBpedia) have a result in LOTUS. We observed little increase of recall on the 'Local Monuments' dataset. While many of the results still come from generic knowledge sources such as DBpedia, the different use cases show that this proportion differs per domain, opening up new perspectives and challenges for application areas, such as Named Entity Disambiguation and Information Retrieval.

In the current version of LOTUS we focus on context-free[31] ranking of results and demonstrate the versatility of LOTUS by measuring its performance and showing how the ranking algorithms affect the search results. A context-dependent ranking mechanism could make use of additional context coming from the query in order to re-score and improve the order of the results. To some extent, context-dependent functionality could be built into LOTUS. However, graph-wide integration with structured data would require a different approach, potentially based on a full text-enabled triplestore (e.g., Virtuoso).

Although further optimization is always possible, the current version of LOTUS performs indexing and querying in an efficient and scalable manner, largely thanks to the underlying distributed architecture. Since the accuracy of LOTUS is case-dependent, future work will evaluate the precision and recall of LOTUS on concrete applications, such as Entity Linking and Network Analysis. The usability of LOTUS could also be compared to standard WWW web search engines, by restricting their results to *filetype:rdf*.

Answer to RQ4 This chapter focused on the fourth research question from this thesis, regarding the enrichment of knowledge on long-tail entities beyond DBpedia. To get much richer and more diverse knowledge on entities that are

31 By "context-free", we mean that the retrieval process cannot be directly influenced by additional restrictions or related information.

underrepresented in DBpedia and similar Wikipedia-based knowledge bases we turned to the Linked Open Data cloud, a collection of tens of billions of statements. We complemented a centralized collection of LOD statements (LOD Laundromat) with LOTUS, an adaptive, centralized, and scalable textual entry point. LOTUS was designed with concrete requirements in mind, such as accessibility, scalability, and diversity of use cases. We showed that LOTUS can search in billions of statements within a second.

The initial experiments with entities from three different entity linking datasets show the potential of LOTUS for entity linking and retrieval of long-tail entities. The fact that candidates for (almost) all long-tail journals and up to 71% of the AIDA-YAGO2 NILs can be potentially found in LOTUS encourages further research in this direction. Future work will investigate how to design, implement, and evaluate a web-of-data-wide entity linking system. This is an unexplored terrain and brings many novel challenges, concerning the efficiency of web-of-data search and a fair evaluation across many heterogeneous knowledge sources. Moreover, we expect that besides the recall, the ambiguity will increase dramatically as well - it is unclear how to reach high precision in entity linking under circumstances of extreme ambiguity. LOTUS and the entire LOD Lab infrastructure are expected to grow and be tuned together with its applications.

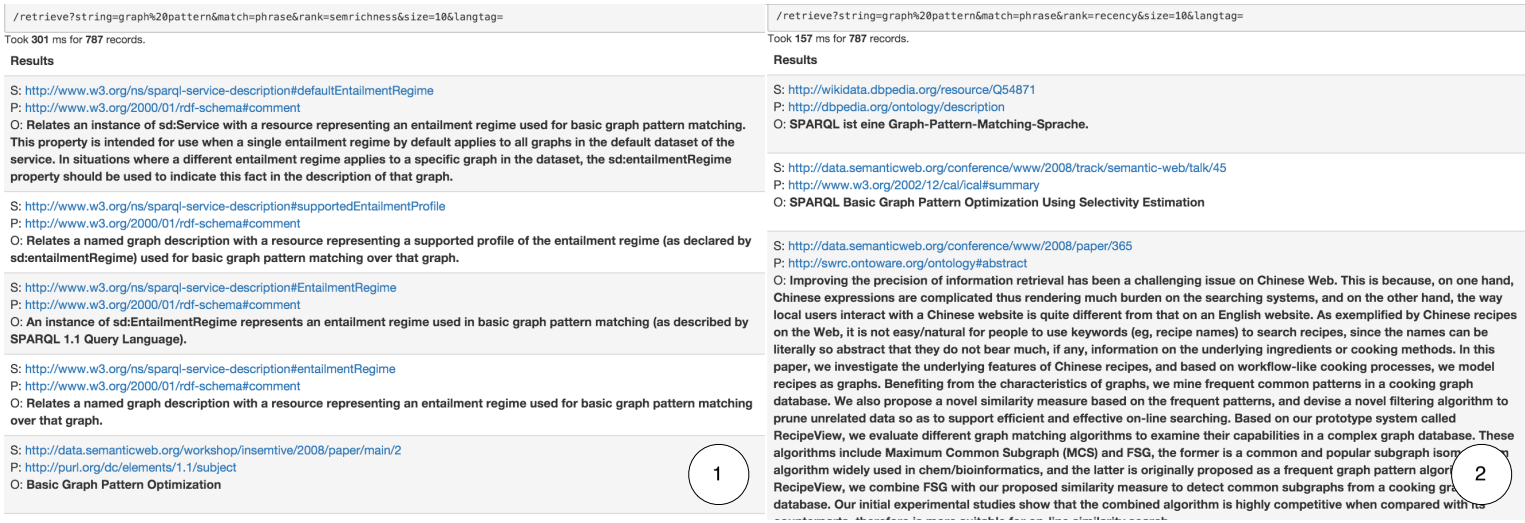


Figure 30: Results of query "graph pattern" with terms-based matching and different rankings: 1) Semantic richness, 2) Recency.

6 THE ROLE OF KNOWLEDGE IN ESTABLISHING IDENTITY OF LONG-TAIL ENTITIES

The NIL (E3) entities do not have an accessible representation, which means that their identity cannot be established through traditional disambiguation. As a consequence, these entities have received little attention in entity linking systems and tasks so far. Similarly, we excluded them from the analysis in chapter 2, given that a large portion of the study would be nonsensical for NIL entities. Nevertheless, the E3 entities can be seen as an extreme variant of the tail (E2) entities, given the non-redundancy of knowledge on these entities, the lack of frequency priors, their potentially extreme ambiguity, and numerousness. Extrapolating the discussion and our findings in previous chapters, we believe that this special class of entities poses a great challenge for state-of-the-art EL systems.

The identity of E3 entities is therefore the main focus of this chapter. Are E3 entities qualitatively different from E1 and E2 entities? What can be done to establish their identity and potentially link to them at a later point? What kind of knowledge can be applied to establish the identity of NILs? How to capture implicit knowledge and fill knowledge gaps in communication? These matters will be considered in this chapter. The main research question that we will investigate is RQ5: *What is the added value of background knowledge models when establishing the identity of NIL entities?*

Due to the unavailability of instance-level knowledge, in this chapter we propose to enrich the locally extracted information with models that rely on background knowledge in Wikidata. This background knowledge is in a structured form and resembles that discussed in the previous chapter, which means that in theory such models could easily be built on top of DBpedia or any other (collection of) LOD knowledge base(s) that LOTUS enables access to. We test our approach on variations of the evaluation data described in chapter 4.

The content of this chapter is a combination of new research and the content published in the following two publications:

1. Filip Ilievski, Piek Vossen, and Marieke van Erp (2017). "Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking." In: *International Conference on Language, Data and Knowledge*. Springer, Cham, pp. 143–149 In this paper I was the main contributor, organized the related work, and provided an argumentation with a set of recommendations.

2. Filip Ilievski, Eduard Hovy, Qizhe Xie, and Piek Vossen (2018). "The Profiling Machine: Active Generalization over Knowledge." In: *ArXiv e-prints*. arXiv: `1810.00782 [cs.AI]` In this preprint, I was the main contributor. I

investigated the idea of building profiles as neural background knowledge models inferred from instance-level data, and tested their accuracy against instance data and human judgments.

6.1 INTRODUCTION

Knowledge scarcity is (unfortunately) a rather prevalent phenomenon, with most instances being part of the Zipfian long tail. Applications in NLP suffer from **hunger for knowledge**, i.e., a lack of information on the tail instances in knowledge bases and in communication. Knowledge missing during NLP system processing is traditionally injected from knowledge bases, which attempt to mimic the extratextual knowledge possessed and applied by humans. However, current knowledge bases are notoriously sparse (Dong et al., 2014), especially on long-tail instances. Not only there are many instances with scarce knowledge (E2 entities), but most (real or imaginary) entities in our world have no accessible representation at all (E3 entities). This poses a limitation to the entity linking task, as we are unable to establish the identity of the vast majority of entities through traditional disambiguation, as defined in the task of entity linking.

Within entity linking, the forms that refer to non-represented entities are simply resolved with a reference to a 'NIL'. The TAC-KBP NIL clustering task takes a step further, as it requires the forms which refer to the same NIL entity within a dataset to be clustered together. We note, however, that the utility of this clustering is limited to the current dataset - provided that no additional information is stored about this entity, it is simply not possible to distinguish it from any other NIL entity found in another dataset. It is also not possible to link any form outside of this dataset to that instance, as it contains no description whatsoever.

In a realistic setting the clustering itself is very complex. Imagine a document D_1 that mentions a NIL entity with a name 'John Smith'. Given another mention with the same form in a document D_2, we need to decide whether the two documents report about the same person, or another person sharing the name. In reality, there are thousands of such documents, most of which describe a different person, but some of which could still be about the same person as D_1. Moreover, there is no guarantee that the information about this entity will be consistent across documents. The lack of frequency priors among these instances (I), the scarcity and non-redundancy of knowledge (II), and the unknown, but potentially extreme, ambiguity (III) make the NIL entities substantially different from the head entities. Considering these factors, the knowledge about the NIL entities needs to be carefully extracted (both high precision and recall), combined, and stored, in order to allow further reasoning.

In this chapter, we investigate **the role of knowledge when establishing the identity of NIL entities mentioned in text**. We expect that even with perfect attribute extraction, it is not always trivial to establish the identity of a long-tail entity across documents due to heterogeneity of information. Namely, the information about an entity found in different documents is not necessarily identical, as these have been written independently of each other and potentially describe

different information relevant in that context. Such knowledge gaps in communication are easily filled by people based on knowledge about associations among facet values, and give rise to a continuously evolving and changing collection of cognitive expectations and stereotypical profiles.[1] People assume that they are entitled to fill knowledge gaps with their expectations unless contradictory evidence is explicitly presented.

Growing amounts of data and the advent of workable neural (deep) models raise the natural question: how can one build models that capture such prominent cognitive skills of people? How can one fill knowledge gaps when these cannot be distilled directly from communication nor retrieved from existing knowledge bases? A popular computational task of completing missing values in a knowledge graph is Knowledge Base Completion (KBC), where the system is asked to add new concrete instance-level facts given other instantial information. In this chapter, we motivate a variant of KBC called **profiling**, where unlike in regular KBC, we predict expectations over value classes rather than predicting specific values with high precision. These expectations should of course be conditioned on whatever (partial) information is provided for any test case, and automatically adjusted when any additional information is provided. We see profiling as a common and potentially more helpful capability for filling gaps in NLP, where knowing preferences or ranges of expectations (rather than concrete values) is often necessary in order to perform reasoning and to direct interpretation on partial and underspecified (long-tail) data.

We design and implement two profiling machines, i.e., background knowledge models that fill knowledge gaps, based on state-of-the-art neural methods. The profiling machines are evaluated extrinsically on a NIL clustering task, where the evaluation data concerns the event participants (of type *Person*) from the referential quantification task ("SemEval-2018 task 5") described in chapter 4. Within this data, we focus on the gun violence domain. The entities in SemEval-2018 task 5 satisfy the properties of long-tail entities described above: they have no representation in the LOD cloud, we can find very little, often inconsistent, information about them in texts, and their ambiguity is potentially very large.

To test the applicability of our profilers to the NIL clustering task, we match the local, incomplete context extracted from text to its corresponding profile. We investigate the interplay between profiles and various intratextual knowledge, distilled with both perfect and automatic extractors. To understand the robustness of our approach, we systematically increase the ambiguity of forms, and investigate how the behavior of our system changes. By combining explicit and implicit knowledge, we are able to gain insight into the role of background knowledge models when establishing identity of long-tail entities. In addition, we perform an intrinsic evaluation of the behavior and performance of the profiling models.

We summarize the contributions of this chapter as follows:

1 We consider the terms 'attribute', 'property', and 'facet' to be synonymous in the context of this work and hence we will use them interchangeably in this chapter.

1. We formulate a set of hypotheses about the role of background knowledge models in establishing identity of long-tail entities **(Section 6.3)**.
2. We motivate a KBC-variant of profiling **(Section 6.4)** to support the interpretation of contextual, long-tail instances in NLP. In profiling, we predict expectations over value classes rather than predicting specific values with high precision.
3. We describe two generic state-of-the-art neural methods that can be easily instantiated into profiling machines for generation of expectations, and applied to any kind of knowledge to fill gaps **(Section 6.4)**.
4. We illustrate the usefulness of the profiling methods *extrinsically* on the referential NLP challenge of establishing identity of long-tail entities **(Section 6.6)**. By establishing identity of entities without popularity/frequency prior and under extremely high ambiguity levels, we are the first to deliberately address the true challenges for the entities in the distributional 'tail'. For the purpose of this evaluation, we adapted the gun violence domain data of SemEval-2018 task 5 (Postma et al., 2018), whose extraction and preparation was detailed in chapter 4.
5. We evaluate the profiling methods *intrinsically* against known values in Wikidata and against human/crowd judgments, and compare the results to gain insight in the nature of knowledge captured by profiles **(Section 6.7)**.
6. We make all code and data available to facilitate future research.

6.2 RELATED WORK

The task of establishing identity of NIL entities to which we attend in this chapter has been motivated by NIL clustering, a recent addition to the standard task of entity linking. We review previous approaches for NIL clustering in section 6.2.1. A knowledge-based approach towards representing identity from text has similarities with previous work on attribute extraction and slot filling, which we review in section 6.2.2. The idea of filling knowledge gaps through profiling default values has similarities with the KRR research on knowledge completion and defaults, as well as other disciplines - these tasks and their typical state-of-the-art approaches are covered in sections 6.2.3 and 6.2.4.

6.2.1 *Entity Linking and NIL clustering*

Entity linking facilitates the integration and reuse of knowledge from existing knowledge bases into corpora, established through a link from an entity mention in text to its representation in a knowledge base. Many entity linking systems have been proposed in the recent years: AGDISTIS (Moussallem et al., 2017), Babelfy (Moro et al., 2014), and DBpedia Spotlight (Daiber et al., 2013), inter alia.

A special category in the EL task are the NIL entities. **NIL entities** are entities without a representation in a knowledge base (Ji and Grishman, 2011). These are typically considered to have low frequencies within a corpus and/or to be domain-specific. Esquivel et al. (2017) report that around 50% of the people mentioned in news articles are not present in Wikipedia. Considering that Wikipedia and its structured data counterparts are almost exclusively used as an anchor in EL, this means that for half of all people mentions, the EL task is nonsensical.

How does one decide between an existing entity representation and a NIL value? EL systems typically use a certain threshold value of the linking confidence for this purpose.

The NIST Text Analysis Conference's Knowledge Base Population (TAC-KBP) (Ji and Grishman, 2011; Ji et al., 2015) has introduced a task of NIL clustering, asking systems to cluster NIL mentions that are coreferential. By doing so, they acknowledge the prominence and the relevance of the NIL entities in corpora, and take a step further towards meaningful interpretation of these mentions. We note, however, that the utility of this outcome is limited to the current dataset - provided that no additional information is stored about this entity, it is simply not possible to distinguish it from any other NIL entity found in another dataset. It is also not possible to link any form outside of this dataset to that instance, as it contains no description whatsoever.

Although state-of-the-art EL systems do not perform NIL clustering or do not detail their clustering algorithm, several NIL clustering approaches were proposed within the TAC EDL competition. Radford et al. (2011) propose three baseline techniques to cluster NIL entities: 1. 'term' (cluster terms that share the same mention) 2. 'coref' (cluster based on in-document coreference) 3. 'KB' (use information from a KB where this NIL actually does exist). Graus et al. (2012) first translate the full documents to TF.IDF weighted vectors, then perform hierarchical agglomerative clustering algorithm on the vectors of all documents that were previously labeled as NIL by the system. Monahan et al. (2011) perform NIL clustering together with EL, by first clustering mentions based on their term and certain features: type, verbal context, etc., and then optionally linking the cluster to a KB entity. When resolving a nominal mention, the most effective approach is to apply within-document coreference resolution to resolve it to a name mention (Ji et al., 2015). Although state-of-the-art within-document coreference resolution performance for nominal mentions is still quite low, linking each identified person nominal mention to its closest person name mention can yield 67% accuracy (Hong et al., 2015).

We observe that the NIL clustering is a fairly marginal part of the overal task of interpretation of entity mentions. The clustering of NILs is mostly either not done at all, or not reported/assumed to be done in a certain default (typically term-based) manner. When reported (typically not most recent work), the method is typically based on the term itself, coreference, and verbal context. In this work, we focus exclusively on establishing identity of NIL entities, and we apply a method that is based on background knowledge models and reasoning over entity attributes.

The issue of knowledge scarcity on long-tail entities is encountered in several approaches which perform entity linking in order to improve the retrieval of information in search engines. Graus et al. (2016) propose to combine dynamic entity representations from various sources in order to improve the entity ranking, whereas Meij et al. (2011) map the search engine queries to DBpedia and propose to extend the set of Linked Open Data sources beyond DBpedia in future work. Although these approaches address a different entity-centric task, within the field of IR, the encountered challenges and the proposed ideas relate to ours.

6.2.2 *Attribute extraction*

Previous work on attribute extraction in the field of Information Retrieval (IR) (Nagy and Farkas, 2012; Zhong et al., 2016) resembles our task and method in several aspects: 1. multiple documents may point to the same person, and there is ambiguity of person names; 2. many entities belong to the long tail; 3. the method is to represent people with attributes that are extracted from text; 4. modeling of instances is based on a restricted set of attributes. However, given that the goal is to find personal websites in a search engine, the context of disambiguation is poorer. Moreover, in most cases, clustering on a document level is sufficient, since there is a single central entity per document.

Our attribute extraction task also resembles the task of slot filling, where the goal is to search the document collection to fill in values for specific attributes ('slots') for given entities. The per-attribute F1-scores show large variance between attributes, between systems, and across datasets. Angeli et al. (2014) reports an average slot filling F1-score of 37%, the average score in (Adel et al., 2016) is 53%, whereas the average F1-scores in the TAC KBP slot filling competitions in 2016 and 2017 for English are between 10% and 25%. Considering this instability in performance and the fact that some of the properties we use were customized to fit the information found in a document, we opted to build our own lexical attribute extractors (described in Section 6.5.4). We report that our performance on the overlapping attributes is comparable to that reported in the aforementioned papers.

6.2.3 *Knowledge Base Completion (KBC)*

Most facet values in both Freebase and Wikidata are missing (Dong et al., 2014). KBC adds new facts to knowledge bases/graphs based on existing ones. Two related tasks are link prediction (with the goal to predict the subject or object of a given triplet, usually within the top 10 results) and triplet completion (a binary task judging whether a given triplet is correct).

In the past decade, KBC predominantly employed deep embedding-based approaches, which can be roughly divided into tensor factorization and neural network models (Ji et al., 2016). TransE (Bordes et al., 2013) and ITransF (Xie et al., 2017) are examples of neural approaches that model the entities and rela-

tions in embedding spaces, and use vector projections across planes to complete missing knowledge. Tensor factorization methods like (Guu et al., 2015) regard a knowledge graph as a three-way adjacency tensor. Most similar to our work, Neural Tensor Networks (Socher et al., 2013) also: 1. aim to fill missing values to mimic people's knowledge; 2. evaluate on structured relations about people; 3. rely on embeddings to abstract from the actual people to profile information.

Universal Schema (US) (Riedel et al., 2013) includes relations extracted from text automatically, which results in larger, but less reliable, initial set of relations and facts. US was designed for the needs of NLP tasks such as fine-grained entity typing (Yao et al., 2013).

As apparent from the amount and diversity of work described here, KBC research is a well-established research direction that encapsulates various efforts to complement knowledge about real-world instances with probable new facts. We define profiling as a variant of KBC that aims: 1. to generate an expectation class for *every facet* of a category/group, rather than suggesting missing facts; 2. to provide a *typical distribution* (not a *specific* value) for the attributes of a specific group. These differences make profiling more useful for reasoning over incomplete data in NLP and AI applications, and related to cognitive work on stereotypes.

KnowledgeVault (KV) (Dong et al., 2014) is a probabilistic knowledge base by Google, which fuses priors about each entity with evidence about it extracted from text. Despite using a different method, the priors in KV serve the same purpose as our profiles: they provide expectations for all unknown properties of an instance, learned from factual data on existing instances. Unfortunately, the code, the experiments, and the output of this work are not publicly available, thus preventing further analysis of these priors, their relation to cognitive/cultural profiling as done by humans, and their applicability in NLP identity tasks.

6.2.4 *Other knowledge completion variants*

Several other, quite distinct research areas, are relevant to profiling. We briefly list some of the most relevant work.

Data imputation. Data imputation refers to the procedure of filling in missing values in databases. In its simplest form, this procedure can be performed by mean imputation and hot-deck imputation (Lakshminarayan et al., 1999). Model-based methods are often based on regression techniques or likelihood maximization (Pearson, 2006). Gautam and Ravi combine a neural network with a genetic algorithm, while Aydilek and Arslan cluster missing values with existing, known values. These efforts focus on guessing numeric and interval-valued data, which is a shared property with the related task of guesstimation (Abourbih et al., 2010). In contrast, profiling aims to predict typical *classes* of values. Moreover, it is unclear how to apply data imputation in NLP use cases.

Estimation of property values. Past work in IR attempted to improve the accuracy of retrieval of long-tail entities by estimating their values based on observed head entities. Most similar to profiling, the method by Farid et al. (2016) estimates a property of a long-tail entity based on the community/ies it belongs

to, assuming that each entity's property values are shared with others from the same community. Since the goal of this line of research is to improve the accuracy of retrieval, the generalization performed is rather ad hoc, and the knowledge modeling and management aspects have not been investigated in depth. Moreover, the code, the experiments, and the results of this work are not publicly available for comparison or further investigation of its usefulness for NLP applications.

Social media analysis. Community discovery and profiling in social media is a task that clusters the online users which belong to the same community, typically using embeddings representation (Akbari and Chua, 2017), without explicitly filling in/representing missing property values.

Local models that infer a specific property (e.g., the tweet user's location) based on other known information, such as her social network (Jurgens, 2013) or expressed content (Mahmud et al., 2012), also address data sparsity. These models target specific facets and data types, thus they are not generalizable to others. Similarly to models in KBC, they lack cognitive support and fill specific instance values rather than typical ranges of values or expectations.

Probabilistic models. Prospect Theory (Kahneman and Tversky, 1979) proposes human-inspired cognitive heuristics to improve decision making under uncertainty. The Causal Calculus theory (Pearl, 2009) allows one to model causal inter-facet dependencies in Bayesian networks. Due to its cognitive background and assumed inter-facet causality, profiling is a natural task for such established probabilistic theories to be applied and tested.

Stereotypes. Stereotype creation is enabled by profiling. The practical uses of stereotypes are vast and potentially ethically problematic. For example, Bolukbasi et al. claim that embedding representations of people carry gender stereotypes; they show that the gender bias can be located in a low-dimensional space, and removed when desired. We leave ethical and prescriptive considerations for other venues, and note simply that artificially removing certain kinds of profiling-relevant signals from the data makes embeddings far less useful for application tasks in NLP and IR when evaluated against actual human performance.

6.3 TASK AND HYPOTHESES

In this section, we provide an explanation of the NIL clustering task. In addition, we present our research question and the set of hypotheses that will be investigated in this chapter.

6.3.1 *The NIL clustering task*

NIL entities are those that do not have a representation in the referent knowledge base K to which we are performing entity linking. Given that most world entities are NIL entities, the real-world ambiguity among NIL entities is potentially far larger compared to the ambiguity of the non-NIL entities that constitute

the majority of our EL datasets. In addition, the lack of an existing representation of these instances in a knowledge base and the minimal textual information on them, means that there is very little redundancy of information on the NIL entities. The ambiguity and knowledge sparsity of these entities requires systems to extract information on NIL mentions carefully and with high precision, but also to be able to have the right expectations about the pieces of information that have been deliberately left out.

NIL clustering Similar to the NIL clustering task introduced in NIST Text Analysis Conference's Knowledge Base Population (TAC-KBP), the aim in this research is to cluster NIL mentions that are coreferential. Formally, the set of forms $f_{i,1}, ..., f_{i,n}$ belongs to the same cluster with the set of forms $f_{j,1} ..., f_{j,m}$ if and only if they are coreferential, i.e., they refer to the same entity instance.

6.3.2 *Research question and hypotheses*

Most entities have no accessible representation, which prevents one to establish their identity by means of traditional entity linking. There is, however, information about these entities in textual documents, which can be used as basis to perform clustering or generate a new representation. Moreover, this knowledge found in text can be enhanced with background knowledge, namely: extratextual, circumtextual, or intertextual knowledge.

In this research, we seek to understand the role of various knowledge that is potentially needed in order to establish identity of a NIL entity. It is unclear to which extent the intratextual knowledge, i.e., the knowledge distilled from text, is sufficient to establish identity of long-tail entities. It is also unclear what is the potential of background knowledge to improve the performance on this task, and what is the sensitivity of the achieved performance to varying degrees of data ambiguity. We will investigate these considerations in this chapter.

Our main research question is:

> **(RQ5)** What is the added value of background knowledge models when establishing the identity of NIL entities?

In order to gain insight into this research question, we put forward six hypotheses about the role of the background knowledge models in the task of NIL clustering. These hypotheses are listed in Table 36.

Moreover, we perform *intrinsic investigation* of the profiling behavior of our components, by comparing them to human judgments and existing instances in Wikidata. Our expectations for the intrinsic evaluations are summarized in Table 37.

ID	Hypothesis	Sec
C1	Assuming that the available information in documents is sufficient, perfect attribute extraction would perform NIL clustering reliably.	6.6.1
C2	Lower-quality (automatic) attribute extraction leads to a decline in the clustering performance compared to perfect extraction.	6.6.1
C3	Assuming sufficient information and perfect attribute extraction, the role of profiling is minor.	6.6.2
C4	Profiling can improve clustering performance when attribute extraction is less accurate.	6.6.2
C5	The overall clustering performance is inversely proportional with ambiguity.	6.6.3
C6	The effect of profiling is larger when the ambiguity is higher.	6.6.3

Table 36: Hypotheses on the role of profiling for establishing identity of long-tail entities.

ID	Hypothesis	Section
P1	Profiling corresponds to the factual instance data.	6.7.1
P2	Profiling corresponds to human expectations.	6.7.2
P3	Profiling is more helpful when the entropy is higher.	6.7.1, 6.7.2
P4	Profiling is more helpful when the value space is larger.	6.7.1
P5	Profiling performance is higher when we know more attributes.	6.7.1

Table 37: Hypotheses on the behavior of profiling.

6.4 PROFILING

6.4.1 *Aspects of profiles*

Following (Ashmore and Del Boca, 1981) we define a profile as a set of beliefs about the attributes of members of some group. A stereotypical profile is a type of *schema*, an organized knowledge structure that is built from experience and carries predictive relations, thus providing a theory about how the social world operates (Kunda, 1999). As a *fast* cognitive process, profiling gives basis for acting in uncertain/unforeseen circumstances (Dijker and Koomen, 1996). Profiles are "shortcuts to thinking", that provide people with rich and distinctive information about unseen individuals. Moreover, they are *efficient*, thus preserving our cognitive resources for more pressing concerns.

Profile accuracy reflects the extent to which beliefs about groups correspond to the actual characteristics of those groups (Jussim et al., 2015). Consensual ones have been empirically shown to be highly accurate, especially the demographic

(race, gender, ethnicity, etc.) and other societal profiles (like occupations or education), and somewhat less the political ones (Jussim et al., 2015). This high accuracy does not mean that profiles will correctly describe any group individual; they are a statistical phenomenon.[2] In that sense, the findings by Jussim et al. that most profiles are justified empirically are of great importance for AI machines: it means that they can be reliably inferred from individual facts, which (unlike many profiles themselves) are readily available.

Profiles exist at various levels of specificity for facets and their combinations. A profile of 20th century French people differs from a profile of 20th century people in general, with more specificity in what kind of food they eat and what movies they watch, or from the profile of French people across all ages. Added information usually causes the initial expectations to change ("shift"), gradually narrowing the denotation group in a transition toward ultimately an individual. The shift may be to a more accurate profile (what in (Stone et al., 1957) is called an "accurately shifted item"), or the opposite ("inaccurately shifted item").

6.4.2 *Examples*

People have no trouble making and using profiles/defaults:

P1 is male and his native language is Inuktitut. What are his citizenship, political party, and religion? Would knowing that he was born in the 19th century change one's guesses?

P2 is a member of the American Senate. Where did he get his degree, and what is his native language?

P3 is an army general based in London. What is P3's stereotypical gender and nationality? If P3 gets an award "Dame Commander of the Order of the British Empire", which expectations would change?

Presumable answers to the above questions are as follows. P1 is a citizen of Canada, votes for the Conservative Party of Canada, and is Catholic. However, P1's 19th century equivalent belongs to a different party. P2 speaks English as main language and graduated at Harvard or Yale University. Finally, the initial expectation on P3 of a male Englishman switches to a female after the award evidence is revealed.

Most of us would agree with the suggested answers. Why/how!? What is it about the Inuktitut language that associates it to the Conservative Party? Why is the expectation about the same person in different time periods different? Why does the sole change of political party change the expectation of one's work position or cause of death? Despite the vast set of possibilities, these kinds of knowledge gaps are easily filled by people based on knowledge about associations among facet values, and give rise to a continuously evolving and changing collection of cognitive expectations and stereotypical profiles.

2 This is a key difference with KBC. KBC looks for the *correct* facet value for an individual, while profiling for the *typical* one.

NLP systems require such human-like expectations in order to deal with knowledge sparsity and the ambiguity of language. In this chapter, we illustrate this on a task of establishing identity of long-tail entities.

6.4.3 *Definition of a profile*

An ideal representation of a long-tail entity would entail combining explicitly stated information in text with implicit expectations based on background knowledge. To illustrate this, let's consider an entity that is described in text as a white resident of Texas. Our implicit knowledge will then enhance this information with expectations about his native language (almost certainly English), religion (e.g., 98% chance of being Christian, 2% others), and gender (e.g., 70% male, 30% female). Here knowledge plays a role both in the local, textual context, as well as in the enrichment of this local context with expectations stemming from a profile. The hunger for knowledge on the long-tail entities can best be bridged by rich knowledge on both sides of the puzzle.

In this work, we use property-value pairs to represent the collection of facts in the local context, which are based on intratextual knowledge found in text. The profiles are formalized as learned probability distributions over the properties with no explicitly mentioned value, and rely exclusively on extratextual knowledge. Generally speaking, the local context and the profiles could be represented in a variety of ways, and could be based on different kinds of knowledge. Alternative representations (e.g., word embeddings) and an integration of additional knowledge (e.g., circumtextual) would be relevant to investigate in future research.

For a set of identical surface forms f mentioned in text, we define its locally learned description, i.e., **local context** *lc(f)*, as a set of property-value pairs extracted from text:

$$lc(f) = \{(p_i, v_{ij}) | p_i \in P \wedge v_{ij} \in V_i\}$$

Here P is the set of the considered properties, and V_i is the domain of values for a property p_i.

Local contexts may not be sufficient to establish identity because: 1. The same form can refer to different local contexts, which might or might not be identical (ambiguity) 2. Different forms sometimes refer to the same local context (variance). Then our task is to establish identity between some, but not all, local contexts. This is non-trivial, since the local concepts are often not directly comparable. For example, while $lc_1 = \{('Birthplace', 'Mexico')\}$ and $lc_2 = \{('Birthplace', 'NewYork')\}$ are certain to be non-identical, none of them is exclusive with $lc_3 = ('Ethnicity', 'Hispanic/Latin')$.

For this purpose, we introduce the notion of **profiles**: globally learned descriptions that help to distinguish locally learned descriptions. Given a set P of properties $p_1, ..., p_N$ and a local context $lc(f) = (p_1, v_{1k}), ..., (p_i, v_{ik})$, we define its profile as a distribution of expected values for the remaining $(N - i)$ properties, namely:

$$\mathtt{profile}(f, P, V, T, K) = lc(f) \cup \{(p_{i+1}, d_{i+1}), ..., (p_N, d_N)\}$$

where $d_{i+1}, ..., d_N$ are distributions of expected values for the properties $p_{i+1}, ..., p_N$ given the known property-value pairs in the local concept $lc(f)$. Besides the form and its local context, the profile of a form depends on: a set of properties P, their corresponding domain of values V, the set of textual documents T, and the background knowledge K. For a local entity context extracted from text, a profile is chosen to be optimal when its property-value pairs have the highest probability given the background knowledge used for training.

Such global profiles would provide us with a global network to compare and disambiguate local contexts that are not directly comparable. By doing so, we expect that we can cluster identical local contexts and pull apart non-identical local contexts with a higher accuracy. We consider all local contexts that share the same profile to be identical, constituting an **equivalence class**.

In order to fulfill their function of successfully disambiguating two locally derived contexts (from text), the set of properties and values constituting a profile needs to fulfill several criteria:

1. to be generic enough. By 'generic', we mean that the properties and their values should be easily applicable to unseen instances, but also to be meaningful to define their identity. The values for each property, furthermore, should be disjoint and easily distinguishable.
2. to be restricted by what can be learned from background knowledge
3. to be based on what can be found in text
4. to be the minimal set of such properties

These criteria are application-dependent: while in information extraction tasks they are driven by the information found in text, an IR application would need to consider criteria like demand/usage, which could, for instance, be captured by query logs (Hopkinson et al., 2018).

6.4.4 *Neural methods for profiling*

In this section we describe two neural architectures for computing profiles at large scale, and baselines for comparison. The implementation code of these architectures and the code to prepare data to train them are available on GitHub: `https://github.com/cltl/Profiling`.

6.4.4.1 *AutoEncoder (AE)*

An autoencoder is a neural network that is trained to copy its input to its output (Goodfellow et al., 2016). Our AE contains a single densely connected hidden layer. Its input $\boldsymbol{x}$ consists of n discrete facets $x = x_1, ..., x_n$, where each x_i is encoded with an embedding of size N_e. For example, if the input is the entity

Angela Merkel, we concatenate n embeddings for its n individual features (nationality: German, political party: CDU, etc.). The total size of the embedding input layer is then $|\boldsymbol{x}| = n * N_e$.

We denote the corresponding output for an input sequence $\boldsymbol{x}$ with $\boldsymbol{z} = g(f(\boldsymbol{x}))$, where f is the encoding and g is the decoding function. The output layer of the AE assigns probabilities to the entire vocabulary for each of the n features. The size of the output layer is a sum over the variable vocabulary sizes of the individual inputs v_i: $|\boldsymbol{z}| = \sum_{i=1}^{n} v_i$.

The AE aims to maximize the probability of the correct class for each feature given inputs $\boldsymbol{x}$, i.e., it is trained to minimize the cross-entropy loss L that measures the discrepancy between the output and the input sequence:

$$L(\boldsymbol{x}, \boldsymbol{z}) = -\sum_{i=1}^{n} [x_i \log z_i + (1 - x_i) \log(1 - z_i)]$$

Due to the sparse input, it is crucial that AE can handle missing values. We aid this in two ways: 1. we deterministically mask all non-existing input values; 2. we apply a probabilistic dropout on the input layer, i.e., in each iteration we randomly remove a subset of the inputs (any existing input is dropped with a probability p) and train the model to still predict these correctly. Although we apply the dropout method to the input layer rather than the hidden layer, we share the motivation with Srivastava et al. (2014) to make the model robust and able to capture subtle dependencies between the inputs. Such dropout helps the AE reconstruct missing data.

6.4.4.2 *Embedding-based Neural Predictor (EMB)*

In our second architecture each input is a single embedding vector **e** rather than a concatenation of n facet-embeddings. For example, the input for the entity Angela Merkel is its fully-trained entity embedding. The size of the input is the size of that embedding: $|\boldsymbol{e}| = N_e$. In the current version of EMB we use pre-trained embeddings as inputs, and we fix their values. Future work can investigate the benefits of further training/tuning.[3]

Like the AE, EMB has one densely connected hidden layer. For an input $\boldsymbol{x}$ and its embedding representation $\boldsymbol{e}$, the corresponding output is $\boldsymbol{z} = g(h(\boldsymbol{e}))$. The output layer of the embedding-based predictor has the same format as the one of the AE, and the same cross-entropy loss function $L(\boldsymbol{x},\boldsymbol{z})$.

6.4.4.3 *Baselines*

We evaluate the methods against two baselines:

Most frequent value baseline (MFV) chooses the most frequent value in the training data for each attribute, e.g., since 14.26% of all politicians are based in the USA, MFV's accuracy for profiling politicians' citizenship is 14.26%. This

3 As pre-trained embeddings are often not available or easy/quick to create, sometimes such training is unavoidable.

baseline indicates for which facets and to which extent our methods can learn dependencies that transcend MFV.

Naive Bayes classifier (NB) applies Bayes' theorem with strong independence assumptions between the features. We represent the inputs for this classifier as one-hot vectors. Naive Bayes classifiers consider the individual contribution of each input to an output class probability. However, the independence assumption prevents it from adequately handling complex inter-feature correlations.

6.4.4.4 *Model Implementation Details*

We experimented with various parameter values suggested in the literature and opted for the following settings. Both neural models use a single dense hidden layer with 128 neurons. For the AE model, we pick an attribute embedding size of $N_e = 30$. These vectors are initialized randomly and trained as part of the network. We set the dropout probability to $p = 0.5$. The inputs to the EMB are 1000-dimensional vectors that were previously trained on Freebase.[4]

Both models were implemented in Theano (Bergstra et al., 2010). We used the ADAM (Kingma and Ba, 2014) optimization algorithm. We train for a maximum of 100 epochs with early stopping after 10 consecutive no-improvement iterations, to select the best model on a held-out validation data. We fix the batch size to 64. When an attribute has no value in an entire minibatch, we apply oversampling: we randomly pick an exemplar that has a value for that attribute from another minibatch and append it to the current one.

6.5 EXPERIMENTAL SETUP

Next, we introduce our experimental setup. We start by discussing a modular end-to-end pipeline for NIL clustering, which will allow us to systematically investigate our research question and hypotheses. Then, we present the data we evaluate on, the evaluation metrics we employ to measure the performance of individual components in the pipeline, and the functionality of these components in detail. The code of all experiments can be found on GitHub: `https://github.com/cltl/LongTailIdentity`.

6.5.1 *End-to-end pipeline*

The testing of our hypotheses C1-C6 is enabled by following different paths in a modular end-to-end architecture for NIL clustering. Figure 31 presents a schematic overview of the components that constitute our end-to-end NIL clustering architecture. The input to the pipeline consists of a set of documents with recognized entity mentions (see section 6.5.2 for details on the data). The goal of the pipeline is to create clusters of NIL entities. The evaluation of these system clusters against the gold clustering output is described in section 6.5.3.

4 These vectors are available at `code.google.com/archive/p/word2vec/`. They correspond and can be mapped to only a subset of all Wikidata entities.

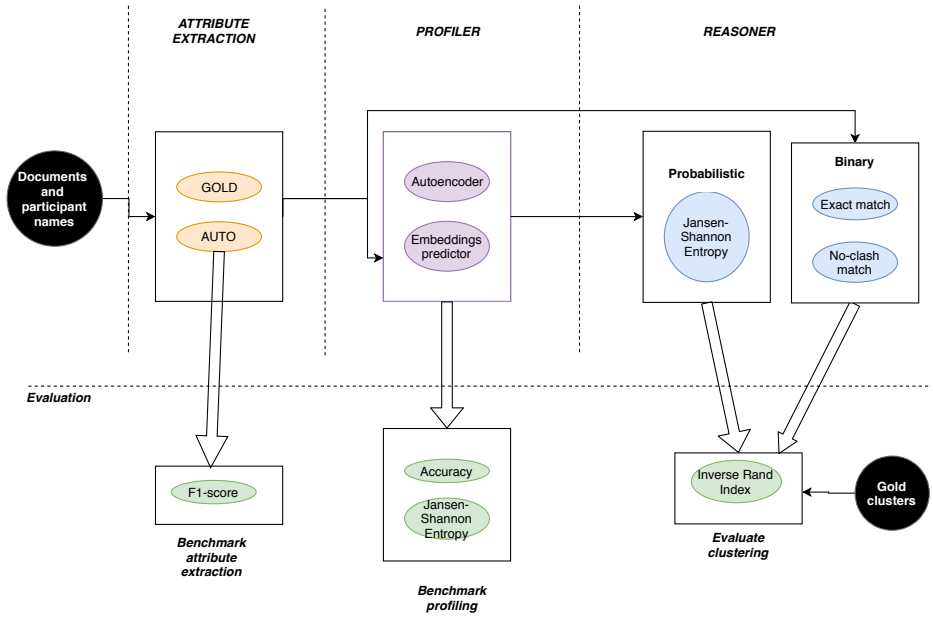


Figure 31: End-to-end architecture for NIL clustering

The pipeline consists of three main processes: attribute extraction, profiler, and reasoner. The attribute extraction process aims to extract explicitly given information about an entity in a document (we refer the reader to section 6.5.4 for more details on our attribute extractors). We experiment with both gold (perfect) attribute extraction, as well as automatic (imperfect) extractors. Once the explicitly mentioned attributes have been extracted, we run profiling models (described in 6.4.4) to obtain default expectations for the attributes which are not explicitly given. In the final third step, we perform reasoning in order to cluster the entity mentions in different documents based on the locally extracted information, potentially enriched by the profiling. Details on the reasoners that we used in this step can be found in section 6.5.5.

6.5.2 *Data*

SemEval-2018 task 5 We test our methods on the Gun Violence domain data from the SemEval-2018 task 5 on 'Counting events and participants in the long tail' (Postma et al., 2018). The creation of this event-based QA task was detailed in chapter 4. Its data consists of documents which potentially describe the same incident; the goal is then to find the number of incidents that satisfy the question constraints, to decide on the documents that report on these incidents, and count the participants in their corresponding roles. Given that this task evaluates the

identity of events, we first need to prepare the data to be suitable for evaluation of the identity of entities.

The SemEval-2018 task 5 data is exceptional in that it describes unknown people participating in local incidents, described in local news documents. As such, these people can safely be assumed to have no representation in the customary knowledge bases, such as DBpedia, but neither in the Linked Open Data cloud. The only way in which we can model the identity of these people is through extensive usage of knowledge, found in the current or other documents, as well as reasoning over conceptual world knowledge.

Each incident comes with structured data describing its location, time, and the age, gender, and name for all participants. From this information, we gather the following properties for each participant: name, gender, age, death year, and death place (the last two attributes apply only if the person was killed).

Data partitions We define two data partitions over this data:

FULL The entire set of 2,261 incidents used in this task comprises our 'full' data partition. This partition contains consistent annotations on an incident level for the attributes: name, age, gender, death year, and death place, for all participants. We do not know whether these attributes are reported in each of the supporting documents - for the purpose of our experiments we assume this is the case, as the structured data was constructed from the news.

PARTIAL The partial data consists of 260 incidents described in 457 documents, capturing 472 participants with 456 distinct names.[5] For a subset of all incidents, we additionally annotated evidence per document and extracted values for 9 additional attributes. This annotation is described in detail next.

Table 38 presents the number of incidents, documents, and participants in each of the two datasets.

Partition	# inc	# docs	# participants
full	2,261	4,479	5,408
partial	260	457	472

Table 38: Number of incidents, documents, and participants in each of the two partitions.

Annotation of explicit values For the partial dataset, we present guidelines[6] to enrich the properties for each entity with 9 additional properties occasionally mentioned in text, namely: Residence, Cause of death, Past convictions/charges, Ethnic group (ethnicity), Education level, Birth place, Native language, Political party, and Religion. For each of these properties and the people participating in

5 We start with the documents from the 241 Gun Violence Corpus (Vossen et al., 2018a) incidents that were annotated with event mentions and we enrich them with additional 50 incidents whose participant shares a name with a participant in the original collection. This results in an initial pool of 291 incidents, which is later filtered as described below, leading to 260 incidents in total.

6 https://docs.google.com/document/d/1rgTdrn-tPoJfPI25-5qOioznmj7un9pi-dO1FyV7pCk/edit#

the article, we perform two types of annotation based on the information given in text:

1. **Structured (incident-level) annotation** we complete the profile of each entity as much as we can based on the information found in text. For instance, let's say that the text contains the following: 'Gaby, who only finished high school this summer, is of Chinese origin. ...'. Then, we fill the ethnic group of Gaby to be 'Chinese/Asian' and her education level - 'high school graduate'.

2. **Mention annotation** we mark the evidence for the profile properties as found in text. In the example given here, we would annotate 'finished high school' as an evidence for Gaby's education level and we would mark 'Chinese' to support the profile trait of Gaby being of Asian descent.

A dedicated web-based tool was created to support this annotation, found at: `http://annpro.fii800.lod.labs.vu.nl/`.[7] Two linguistics Master students were hired as annotators for a day per week over a period of three months. The inter-annotator agreement for the structured annotation is 0.852, whereas the agreement on document- and sentence-level marking of evidence is 0.849 and 0.648, correspondingly.

The additional annotation performed by our students was only performed on the documents and incidents that belong to the 'partial' dataset. For the full dataset, we only use the properties provided by the original data source.

Postprocessing In a postprocessing step, we remove: 1. documents and incidents that were disqualified by our annotators 2. incidents without new annotation of structured data 3. documents without any annotation 4. participants with no name. In addition, we merged the incidents with different IDs which were identical, as well as the participants that appeared in multiple incidents.

Increase of ambiguity Besides including incidents with participants with the same name, within our experiments we systematically and artificially increase the ambiguity, simply by changing names of people in the structured data.

Combining the four ambiguity levels and two data partitions leads to eight datasets in total. The ambiguity statistics for all eight datasets are presented in Table 39.

6.5.3 *Evaluation*

We evaluate the accuracy of different methods for establishing identity of entities with the clustering metric Adjusted Rand Index (ARI). We compute ARI between a system output and the gold clustering. The score produced by this metric ranges between 0 and 1, where larger values stand for better clustering.

In addition, we perform intrinsic evaluation of the individual components in our end-to-end pipeline. Namely, we benchmark the property extractors by using

7 Source code: `https://github.com/cltl/AnnotatingEntityProfiles`

Partition	Modification	Unique SFs	Mean ambiguity
Partial	Original data	456	1.035
	Same first name 'John'	325	1.452
	Same last name 'Smith'	377	1.252
	Same name 'John Smith'	1	472
Full	Original data	5,329	1.015
	Same first name 'John'	3,547	1.525
	Same last name 'Smith'	3,557	1.520
	Same name 'John Smith'	1	5,408

Table 39: Ambiguity statistics. Total number of unique instances in the partial data: 472, total instances in the full data: 5,408. $Meanambiguity = \#uniqueInstances/\#uniqueSFs$.

the customary metrics of precision, recall, and F1-score. These evaluation results are provided in section 6.5.4.

Regarding the profiling machines, we measure their intrinsic accuracy by evaluating them against 'correct' values in Wikidata. We also measure their correspondence to human expectations through a Jansen-Shannon divergence against the distribution of crowd judgments. We refer the reader to section 6.7 for the results of these analyses.

6.5.4 *Automatic attribute extraction*

For the purposes of our hypotheses, we also seek to understand the usefulness of profiling over both perfect and imperfect attribute extraction. We have thus built the following automatic extraction strategies from text:

1. Proximity algorithm (see Algorithm 1) which assigns spotted phrases in text to their closest mention of person, as long as this occurs in the same sentence. This strategy was applied to all properties, except for the gender.

2. Coreference algorithm (see Algorithm 2) looks for gender keywords in the coreferential phrases to a person.

We have benchmarked the automatic attribute extraction with both strategies against the gold extraction data we possess: see Table 40 for a comparison against known values in the Gun Violence Database (GVDB) (Pavlick et al., 2016) and Table 41 for a comparison against a subset of the SemEval-2018 task 5 data (see section 6.5.2 for more details on this data).

As we discussed in section 6.5.4, it is not trivial to compare the performance of our extractors directly to previous research on attribute extraction or slot filling. Fair comparison is prevented by differences in the test data (domain) and the considered properties. Given that the performance per property varies greatly and

Algorithm 1 Automatic attribute extraction: Proximity strategy

```
Require: attr, text, person_names
regexes ← set_regexes(attr)
for all person_name ∈ person_names do
    person_spans[person_name] ← findall(person_name, text)
    coref_spans[person_name] ← coreference(person_spans, text)
end for
for all regex ∈ regexes do
    matches ← findall(regex, text)
    for all match ∈ matches do
        the_person ← find_closest_person(match,
        person_spans, coref_spans, sentence_boundaries)
        the_person[attr]+ = match
    end for
end for
for all person_name ∈ person_names do
    person_name[attr] ← find_closest(person_name[attr])
end for
return person_names
```

Algorithm 2 Automatic attribute extraction: Coreference strategy

```
Require: attr, text, person_names
regexes ← set_regexes(attr)
for all person_name ∈ person_names do
    person_spans[person_name] ← findall(person_name, text)
    coref_spans[person_name] ← coreference(person_spans, text)
end for
for all person_name ∈ person_names do
    genders[person_name] ← cooccurring(keywords,
    person_name, person_spans, coref_spans)
end for
for all person_name ∈ person_names do
    gender[person_name] = most_common(genders[person])
end for
return gender
```

	precision	recall	F1-score
age	97.81	85.44	91.21
ethnicity/race	40.00	8.70	14.29
gender	81.78	57.14	67.28

Table 40: Benchmarking the attribute extractors on gold data about 1000 articles from the GVDB dataset.

	method	precision	recall	F1-score	#gold	#sys
cause of death	pattern	39.01	17.83	24.47	488	223
past conviction	pattern	9.52	25.00	13.79	32	84
education level	pattern	62.16	19.66	29.87	117	37
ethnicity/race	pattern	0.00	0.00	0.00	10	16
religion	pattern	50.00	11.76	19.05	17	4
age group*	pattern	93.20	25.98	40.63	739	206
gender*	coref	88.03	13.62	23.60	756	117
gender*	pattern	71.85	12.83	21.77	756	135
birthplace	pattern	0.00	0.00	0.00	6	1
residence	pattern	38.40	23.82	29.40	403	250

Table 41: Benchmarking the attribute extractors on the gold data from the 456 documents of the SemEval dataset. For the attributes marked with '*', we do not have mention- or document-level annotation, hence the reported recall (and consequently F1-score) should be considered a lower bound for the 'real' score.

covers almost the entire spectrum of possible scores, it is particularly nonsensical to compare the average F1-scores over all properties, as this would directly be determined by the choice of properties.

Keeping in mind that the data differences persist, we make an attempt to compare individual overlapping properties in order to gain insight into the magnitude of our scores. For this purpose, let us consider them in relation to two previous slot filling systems: (Angeli et al., 2014) (for brevity, we will refer to this system with 'SF1') and (Adel et al., 2016) ('SF2'). Although the majority of the properties we consider have no counterpart in these past works, we report comparisons of the properties that can be matched. The F1-score of extracting state(s) of residence by SF1 and SF2 is 12 and 31, respectively - whereas our performance on the SemEval-2018 task 5 dataset is 29.40; SF1 and SF2 extract age with F1-scores of 93 and 77 compared to our performance of 91.21 (GVDB) and 40.63 (SemEval)[8]; the cause of death F1-scores of these systems are 55 and 31,

8 This F1-score is a lower bound given that the measured recall is lower than the real one, see Table 41. Considering the unreliability of this particular F1-score, it is more accurate to compare the gender attribute on the SemEval dataset against past work in terms of *precision*.

whereas ours is 24.47. This limited evidence tells us that our attribute extractors have comparable performance to that of past work.

Algorithm 3 Probabilistic reasoner based on Jansen-Shannon entropy and Density-Based Spatial Clustering

Require: known_attributes, profiles, person_names, EPS
distances ← ∅
for all name1, name2 ∈ person_names **do**
    attributes[name1] ← known_attributes(name1) ∪ profiles(name1)
    attributes[name2] ← known_attributes(name2) ∪ profiles(name2)
    distances(name1, name2) ← avg_js_entropy(attributes[name1], attributes[name2])
end for
clusters ← DBSCAN_clustering(distances, EPS)
return clusters

6.5.5 *Reasoners*

Once the attributes have been extracted, we need a strategy to decide whether two local instances that share a name are to be clustered or not. We use the following three strategies that compute clusters over the entities with extracted attributes:

1. Exact match (EX) - this reasoner clusters two entities only when all their attribute values are identical, formally:

$$EX : (I_1 = I_2) \iff I_1(p) = I_2(p), \forall p \in P$$

2. No-clash (NC) - this reasoner establishes identity whenever two local representations have no conflicting properties, formally:

$$NC : (I_1 = I_2) \iff \nexists p \in P, I_1(p) \neq I_2(p)$$

   We apply these two reasoners (EX and NC) on the properties extracted from text, but also on an extension of these properties provided by our profiler. In order to achieve the latter, we discretize the probabilities provided by the profiler based on a threshold parameter, τ. Namely, we keep the property values with a probability larger than τ, and discard the others.

3. Probabilistic reasoner (PR) - Since the profiler computes probabilistic distributions over values, we implemented a probabilistic reasoner that clusters entities based on the similarity of their attribute distributions. This reasoner first computes the average pairwise divergence (based on the Jansen-Shannon entropy) between the attribute distributions of all entities that share a name, resulting in a distance matrix.[9] Subsequently, a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et

9 We report results based on the mean divergence over all properties; using maximum instead of mean distance yielded comparable results.

al., 1996) is run over this matrix to obtain clusters. The resulting clusters depend on a threshold called EPS, which determines the maximum distance allowed within a cluster. The functionality of the PR reasoner is described in Algorithm 3.

6.6 EXTRINSIC EVALUATION

In this section we present the results of our experiments on using the profiler within an end-to-end pipeline to tackle the task of NIL clustering. These results provide evidence for the research question that we investigate in this chapter, and give answer to the hypotheses C1 through C6. Namely, section 6.6.1 shows the performance of our pipeline with perfect and imperfect attribute extraction, thus addressing C1 and C2. In section 6.6.2, we provide the results of running the profiler on top of these attribute extractors (C3 and C4). Finally, we analyze the impact of the task ambiguity on the clustering performance (C5) and its relation to the effectiveness of our profiler (C6).

6.6.1 *Using explicit information to establish identity*

In C1, we hypothesized that the performance of clustering by attribute reasoning depends on two factors: availability of information and quality of extraction. An ideal availability of information and perfect extraction would lead to a perfect accuracy on the task of establishing identity of entities.

We first present the clustering performance of various combinations of perfectly extracted attributes in Tables 42 and 43. Namely, besides the name baseline (p0) and the union of all properties, we consider additional ten combinations of properties. We note that the sets p1 through p4 rely on properties annotated in documents by our students; hence, these attributes can safely be assumed to appear in the annotated documents. The combinations p5 through p9 rely on the structured incident data as found on the GVA website. For these sets of properties, we make an assumption that they are consistently mentioned in all reporting documents. As we do not know how often this assumption holds, the obtained scores for p5-p9 should thus be seen as upper bound results for the real scores. Finally, we report performance of p10, which represents the set of properties that were successfully mapped to background knowledge, and will be used for a comparison to the profiler later in this chapter.

This analysis provides insight into the availability and the diversity of information about an entity across documents, as well as into the discriminative power of this information. We observe that correct extraction of the properties leads to almost perfect accuracy on the original data, but notably lower accuracy on the more ambiguous subsets. Hence, assuming perfect attribute extraction, the properties found in text would be mostly sufficient to establish the identity of entities in the original dataset, but this information becomes increasingly insuf-

	PARTIAL				FULL			
	PD1	PD2	PD3	PD4	FD1	FD2	FD3	FD4
p0	0.988	0.413	0.654	/	0.933	0.197	0.236	/
p1	0.846	0.598	0.767	0.006	/	/	/	/
p2	0.698	0.535	0.685	0.018	/	/	/	/
p3	0.679	0.522	0.666	0.018	/	/	/	/
p4	0.679	0.522	0.666	0.018	/	/	/	/
p5	0.993	0.701	0.818	0.005	0.951	0.354	0.358	0.012
p6	0.993	0.821	0.821	0.010	0.975	0.466	0.467	0.021
p7	0.997	0.904	0.932	0.037	0.976	0.472	0.473	0.037
p8	0.993	0.909	0.974	0.086	0.976	0.473	0.473	0.040
p9	0.997	0.916	0.983	0.093	0.976	0.473	0.473	0.040
p10	0.905	0.807	0.873	0.030	0.975	0.466	0.467	0.021
all	0.681	0.636	0.681	0.384	/	/	/	/

Table 42: Clustering accuracy with various combinations of gold properties, using the 'exact' match clustering. Combinations of properties: p0=name baseline, p1=(name, educationlevel, causeofdeath), p2=(name,educationlevel, causeofdeath, residence), p3=(name, educationlevel, causeofdeath, residence, religion, ethnicity, pastconviction), p4=(name, educationlevel, causeofdeath, residence, religion, ethnicity, pastconviction, birthplace), p5=(name, age), p6=(name, age, gender), p7=(name, age, gender, death date), p8=(name, age, gender, death place), p9=(name, age, gender, death place, death date), p10=(name, causeofdeath, religion, ethnicity, pastconviction, age, gender, occupation, nativelanguage, politicalparty), all. Datasets: PD1=original partial data, PD2=partial data with same first name, PD3=partial data with same last name, PD4=partial data with all same names, FD1=original full data, FD2=full data with same first name, FD3=full data with same last name, FD4=full data with all same names. Some cells in the full datasets are empty due to unavailability of information.

ficient as the ambiguity grows.[10] The results also show that certain attributes, such as age and gender, have very large discriminative power, as long as they are consistently reported across documents.

Next, in Table 45, we show the clustering accuracy when automatic attribute extraction is employed. While the no-clash reasoner comes closer to the perfect extraction performance than the exact reasoner, the clustering performance of the automatic attribute extractors is consistently lower than that of the perfect attribute extractors, as expected in our hypothesis C2. This difference in clustering performance grows together with the ambiguity of data. These findings are not

10 We unfortunately do not know if some of the attributes (e.g., gender and age) presented in the gold data occur in each document, hence the scores presented here might be higher than the real ones.

	PARTIAL				FULL			
	PD1	PD2	PD3	PD4	FD1	FD2	FD3	FD4
p0	0.988	0.413	0.654	/	0.933	0.197	0.236	/
p1	0.987	0.611	0.783	0.002	/	/	/	/
p2	0.976	0.634	0.861	0.012	/	/	/	/
p3	0.976	0.632	0.865	0.013	/	/	/	/
p4	0.976	0.632	0.865	0.013	/	/	/	/
p5	0.991	0.687	0.802	0.005	0.936	0.348	0.352	0.012
p6	0.991	0.807	0.804	0.010	0.965	0.459	0.459	0.019
p7	0.995	0.852	0.861	0.034	0.965	0.462	0.462	0.030
p8	0.991	0.863	0.903	0.078	0.965	0.463	0.463	0.034
p9	0.995	0.869	0.913	0.088	0.965	0.463	0.463	0.035
p10	0.991	0.840	0.886	0.025	0.965	0.459	0.459	0.019
all	0.980	0.895	0.932	0.366	/	/	/	/

Table 43: Clustering accuracy with various combinations of gold properties, using the 'no-clash' match clustering. Combinations of properties: p0=name baseline, p1=(name, educationlevel, causeofdeath), p2=(name,educationlevel, causeofdeath, residence), p3=(name, educationlevel, causeofdeath, residence, religion, ethnicity, pastconviction), p4=(name, educationlevel, causeofdeath, residence, religion, ethnicity, pastconviction, birthplace), p5=(name, age), p6=(name, age, gender), p7=(name, age, gender, death date), p8=(name, age, gender, death place), p9=(name, age, gender, death place, death date), p10=(name, causeofdeath, religion, ethnicity, pastconviction, age, gender, occupation, nativelanguage, politicalparty), all. Datasets: PD1=original partial data, PD2=partial data with same first name, PD3=partial data with same last name, PD4=partial data with all same names, FD1=original full data, FD2=full data with same first name, FD3=full data with same last name, FD4=full data with all same names. Some cells in the full datasets are empty due to unavailability of information.

surprising given the benchmark results of these automatic extractors presented in section 6.5.4.

In addition, we observe that both for gold and for automatically extracted information, the improvement in terms of clustering compared to the name baseline is larger when the ambiguity of data is larger.

6.6.2 *Profiling implicit information*

The profiler can be seen as a 'soft' middle ground between the exact reasoner and the no-clash reasoner. The former establishes identity only when all known attributes between two local representations are identical, and the latter - when-

ever two local representations have no conflicting properties. The profiler relies on background knowledge in order to fill the gaps for the unknown properties instead of making a hard decision. We hence expect the profiler to be superior over the above two baselines, as it employs external knowledge to bring closer or further two ambiguous representations. We hypothesized that the role of the profiler is much more important when the attribute extraction is imperfect (hypotheses C3 and C4).

Tables 44 and 45 show the impact of the profiler when applied on top of either perfect or imperfect attribute extraction.[11] We note that the results reported here use the following set of attributes: name, religion, ethnic group, cause of death, gender, occupation, age group, native language, and political party. These properties correspond to the property set *p10* in Tables 42 and 43. Notably, as the properties: native language, occupation, and political party do not occur in text, their values are always filled implicitly by the profiler. The expectations for the other properties are filled by the profiler only when they are not extracted from text.

As discussed previously, in order to experiment with the profiler in combination with the EX and NC reasoners, the probabilities produced by the profiler are first discretized by using a cut-off threshold, τ. Larger values of this threshold mean stricter criteria for inclusion of a certain attribute value; hence, a τ value of 1.0 excludes all output from the profiler, whereas lower τ values allow for more profiler expectations to be included. We also test the usefulness of the profiler in combination with the probabilistic reasoner, PR; in this case, we use the probability distribution as produced by the profiler, in combination with the extracted property values from text. Here, to acknowledge the effect of the clustering parameter of maximum distance (EPS) in the DBSCAN algorithm, we report results with five different EPS values, ranging between 0.05 and 0.5 (higher EPS values lead to less clusters with larger sizes). The number of clusters for different values of the EPS parameter are given in Table 46.

Table 44 shows that enhancing the gold properties by profiling yields comparable results to the ones obtained by the baselines, which corresponds to our hypothesis C3. This observation holds for all three reasoners: exact, no-clash, and probabilistic reasoner. Given that the results of clustering by reasoning over the perfectly extracted properties are relatively high (Tables 42 and 43), there is little room to improve them by profiling. Still, enhancing these gold properties by profiling yields slight improvement over the results of the extractors in several occasions. For instance, combining the no-clash reasoner with profiling at $\tau = 0.90$ (i.e., keeping the profiling values with a probability of 0.90 or higher), leads to the best score on the PD2 dataset. Similarly, the profiler increases the performance on top of the exact baseline for the datasets PD1 and PD2. While for the other datasets the profiler does not improve the baseline performance, we note that its output is fairly robust - even when much of the profiler output is included (e.g., with $\tau = 0.51$), the results do not decline to a large extent.

11 We leave out the results on the full dataset with maximum ambiguity as these are consistently very low.

	PARTIAL				FULL		
	PD1	PD2	PD3	PD4	FD1	FD2	FD3
name baseline (po)	0.988	0.413	0.654	/	0.933	0.197	0.236
exact (EX) reasoner	0.905	0.807	0.873	**0.030**	**0.975**	**0.466**	**0.467**
+ profiler ($\tau = 0.99$)	0.905	0.807	0.873	**0.030**	0.973	0.384	0.378
+ profiler ($\tau = 0.90$)	0.911	0.810	0.870	0.029	0.973	0.384	0.378
+ profiler ($\tau = 0.75$)	0.911	0.810	0.870	0.029	0.949	0.313	0.309
+ profiler ($\tau = 0.51$)	0.910	0.810	0.869	0.029	0.965	0.367	0.360
no-clash (NC) reasoner	**0.991**	0.840	**0.886**	0.025	0.965	0.459	0.459
+ profiler ($\tau = 0.99$)	0.989	0.839	0.885	0.022	0.963	0.369	0.358
+ profiler ($\tau = 0.90$)	0.977	**0.843**	0.857	0.019	0.963	0.371	0.360
+ profiler ($\tau = 0.75$)	0.960	0.836	0.848	0.023	0.934	0.298	0.294
+ profiler ($\tau = 0.51$)	0.925	0.811	0.857	0.020	0.934	0.354	0.347
PR reasoner, EPS=0.01	0.906	0.808	0.871	**0.030**	0.973	0.384	0.378
PR reasoner, EPS=0.05	0.914	0.808	0.871	0.026	0.934	0.290	0.285
PR reasoner, EPS=0.1	0.950	0.811	0.839	0.002	0.934	0.273	0.276
PR reasoner, EPS=0.2	0.987	0.476	0.673	0.000	0.933	0.198	0.236
PR reasoner, EPS=0.5	0.988	0.413	0.654	0.000	0.933	0.197	0.236

Table 44: Inspection of the effect of profiling on the clustering performance, on top of gold attribute extraction. Datasets: PD1=original partial data, PD2=partial data with same first name, PD3=partial data with same last name, PD4=partial data with all same names, FD1=original full data, FD2=full data with same first name, FD3=full data with same last name, FD4=full data with all same names. The set of properties used by the baselines and by the profiler corresponds to p10 in Tables 42 and 43. We report results of applying the profiler in combination with each of the baselines, by first discretizing its probability distribution with a threshold, τ. In addition, we report results of using the probability distributions as provided by the profiler, in combination with Jansen-Shannon entropy and DBSCAN clustering, which corresponds to the probabilistic reasoner (PR) described in section 6.5.5. We vary the clustering coefficient, EPS, between 0.01 and 0.5.

	PARTIAL				FULL		
	PD1	PD2	PD3	PD4	FD1	FD2	FD3
name baseline (po)	**0.988**	0.413	0.654	/	**0.933**	0.197	0.236
exact (EX) reasoner	0.609	0.385	0.534	**0.002**	0.780	0.212	0.237
+ profiler ($\tau = 0.99$)	0.622	0.390	0.545	0.001	0.784	0.211	0.237
+ profiler ($\tau = 0.90$)	0.637	0.398	0.555	0.001	0.787	0.211	0.238
+ profiler ($\tau = 0.75$)	0.666	0.397	0.576	0.001	0.787	0.206	0.231
+ profiler ($\tau = 0.51$)	0.693	0.401	0.576	0.001	0.848	0.215	0.250
no-clash (NC) reasoner	0.944	**0.506**	**0.768**	**0.002**	0.910	0.229	0.258
+ profiler ($\tau = 0.99$)	0.943	0.502	0.745	0.001	0.916	**0.235**	**0.274**
+ profiler ($\tau = 0.90$)	0.932	0.485	0.713	0.001	0.920	0.229	0.271
+ profiler ($\tau = 0.75$)	0.898	0.459	0.657	0.001	0.910	0.225	0.260
+ profiler ($\tau = 0.51$)	0.825	0.410	0.620	0.000	0.897	0.210	0.245
PR reasoner, EPS=0.01	0.645	0.398	0.555	0.001	0.786	0.211	0.238
PR reasoner, EPS=0.05	0.726	0.396	0.595	0.001	0.849	0.210	0.244
PR reasoner, EPS=0.1	0.858	0.384	0.631	0.000	0.901	0.201	0.236
PR reasoner, EPS=0.2	0.974	0.413	0.654	0.000	0931	0.197	0.236
PR reasoner, EPS=0.5	**0.988**	0.413	0.654	0.000	**0.933**	0.197	0.236

Table 45: Inspection of the effect of profiling on the clustering performance, on top of automatic attribute extraction. Datasets: PD1=original partial data, PD2=partial data with same first name, PD3=partial data with same last name, PD4=partial data with all same names, FD1=original full data, FD2=full data with same first name, FD3=full data with same last name, FD4=full data with all same names. The set of properties used by the baselines and by the profiler corresponds to p10 in Tables 42 and 43. We report results of applying the profiler in combination with each of the baselines, by first discretizing its probability distribution with a threshold, τ. In addition, we report results of using the probability distributions as provided by the profiler, in combination with Jansen-Shannon entropy and DBSCAN clustering, which corresponds to the probabilistic reasoner (PR) described in section 6.5.5. We vary the clustering coefficient, EPS, between 0.01 and 0.5.

EPS		PD1	PD2	PD3	PD4	FD1	FD2	FD3
gold	0.01	503	475	489	46	5,344	4,354	4,113
	0.05	498	467	480	30	5,335	4,238	3,932
	0.1	480	439	446	4	5,335	4,220	3,926
	0.2	459	360	390	1	5,329	3,610	3,563
	0.5	456	325	377	1	5,328	3,547	3,557
auto	0.01	608	526	566	31	7,597	5,865	5,954
	0.05	578	475	525	19	6,768	5,027	5,091
	0.1	521	399	451	5	5,967	4,214	4,280
	0.2	464	332	384	1	5,376	3,574	3,600
	0.5	456	325	377	1	5,328	3,547	3,557

Table 46: Number of clusters for different values of the EPS parameter. Datasets: PD1=original partial data, PD2=partial data with same first name, PD3=partial data with same last name, PD4=partial data with all same names, FD1=original full data, FD2=full data with same first name, FD3=full data with same last name, FD4=full data with all same names.

We expect that the profiler has a larger effect when combined with automatic attribute extraction (C_4). While we do not observe significant improvement over the baselines, the results in Table 45 demonstrate that the profiler has certain impact when applied in combination with imperfect attribute extraction. Especially, we observe that the profiler consistently improves the performance on top of the exact reasoner. Furthermore, this improvement becomes larger when more of the profiling values are included, i.e., the improvement on the exact reasoner is inversely proportional with the probability threshold τ. This observation means that the profiler is able to fulfill its role of normalizing the attribute values found in different documents. Namely, as these documents are written independently from each other, certain attribute values are reported only in a subset of them. In addition, the low recall of our automatic extraction tools might increase this inconsistency of information between documents. The results show that the profiler is able to make this reasoner more robust and normalize certain knowledge gaps.

Profiling in combination with the no-clash reasoner yields comparable results to those of only using the automatic extraction. Concretely, profiling improves the performance of the no-clash reasoner for all full datasets (FD1, FD2, and FD3), whereas it decreases the baseline performance on all partial datasets. In general, the profiling results seem to be fairly robust when combined with this reasoner as well.

The probabilistic reasoner also yields certain promising results on the full datasets, whereas its best scores on the partial datasets are obtained for the high-

est value of the EPS distance parameter (0.5), and correspond to the results of the name baseline.

In section 6.5.4, we observed that the performance of our extractors, especially in terms of recall, is relatively low. Consequently, the extracted local contexts from text are largely incomplete, and in some cases contain incorrect values. As the output of the profiler is dependent on the input it receives, this imperfection of the extracted information would directly influence the usefulness of the generated profiles. Namely, a very sparse/incomplete input (caused by low extraction recall) could lead to a profile that represents a more generic group, whereas incorrect input (caused by low precision) might generate a profile that represents an entirely different group of entities. Future work should investigate whether an extended set of attributes and a different kind of am attribute extractor yield similar results for the hypotheses C3 and C4.

6.6.3 *Analysis of ambiguity*

We expect that the clustering performance is counter-proportional to the ambiguity of a dataset (C5), i.e., higher ambiguity leads to lower clustering performance. This is a clear trend that is visible in all result tables. The clustering on the original dataset with minimal ambiguity is close to perfect, whereas the performance of clustering for the datasets with maximum ambiguity is close to zero.

We also hypothesized that the impact of the profiler is higher when there is more ambiguity (C6). Tables 44 and 45 show no clear relation between the usefulness of the profiler and the data ambiguity. Future work should investigate whether this conclusion is confirmed for a larger set of properties.

6.7 INTRINSIC ANALYSIS OF THE PROFILER

In this section, we investigate the intrinsic behavior of the profiler, by comparing it against factual data from Wikidata, as well as against human judgments collected with a crowdsourcing task. We also investigate the relation of the profiling performance to various properties of the data, such as its entropy and value size. These investigations provide evidence for the hypotheses P1-P5.

6.7.1 *Comparison against factual data*

6.7.1.1 *Data*

No existing dataset is directly suitable to evaluate profiling. We therefore chose People, since data is plentiful, people are multifaceted, and it is easy to spot problematic generalizations. We defined three typed datasets: people, politicians, and actors, each with the same stereotypical facets, such as nationality, religion, and political party, that largely correspond to some facets central in social psychology research. We created data tables by extracting facets of people from Wikidata.

Table 47 lists all attributes, each with its number of distinct categories v_i, total non-empty values (n_{ex}), and entropy values (H_i and $H_i\prime$) on the training data.

The goal is to dynamically generate expectations for the same set of 14 facets in each dataset. We evaluate on multiple datasets to test the sensitivity of our models to the number of examples and categories. The largest dataset describes 3.2 million people, followed by the data on politicians and actors, smaller by an order of magnitude. As pre-trained embeddings are only available for a subset of all people in Wikidata (see section 6.4.4.4), we cannot evaluate EMB directly on these sets. Hence, to facilitate a fair comparison of both our models on the same data, we also define smaller data portions for which we have pre-trained embeddings. We randomly split each of the datasets into training, development, and test sets at 80-10-10 ratio.

6.7.1.2 *Quantification of the Data Space*

We quantify aspects of profiling through the set of possible outcomes and its relation to the distribution of values.

The total size of the data value space is $d_{size} = \prod_{i=1}^{n} v_i$, where n is the number of attributes and v_i is the size of the category vocabulary for an attribute x_i (e.g. $v_i = |\{\text{Swiss, Dutch}\dots\}|$ for $x_i = nationality$). We define the **average training density** as the ratio of the total data value size to the overall number of training examples n_{ex}: $d_{avg-d} = d_{size}/n_{ex}$. As an illustration, we note that the full dataset on People has $d_{size} = 10^{39}$ and $d_{avg-d} = 10^{32}$.

For the i-th attribute x_i, the entropy H_i of its values is their 'dispersion' across its v_i different categories. The entropy for each category j of x_i is computed as $-p_{i,j}\log p_{i,j}$, where $p_{i,j} = n_{ex}(i,j)/n_{ex}(i)$. The **entropy** of x_i is then a sum of the individual category entropies: $H_i = -\sum_{j=1}^{v_i} p_{i,j}\log p_{i,j}$, whereas its **normalized entropy** is limited to $[0,1]$: $H_i\prime = H_i/\log_2(n_{ex}(i))$. Entropy is a measure of informativeness: when $H_i\prime = 0$ there is only one value for x_i; when all values are equally spread the entropy is maximal, $H_i\prime = 1$ (with no MFV).

Of course, we do not know the true distribution but only that of the sparse input data. Here we assume our sample is unbiased. Table 47 shows that, e.g., *educated at* consistently has less instance values and a 'flatter' value distribution (= higher $H_i\prime$) than *sex or gender*, where the category *male* is dominant on any dataset, except for actors. The entropy and the categories size together can be seen as an indicator for the relevance of a facet for a dataset, e.g., $H_i\prime$ and v_i of *position held* are notably the lowest for actors. We expect MFV to already perform well on facets with low entropy, whereas higher entropy to require more complex dependencies.

6.7.1.3 *Results*

We evaluate by measuring the correctness of predicted (i.e., top-scoring) attribute values against their (not provided) true values, evaluated only on exemplars that were not included in the training data.

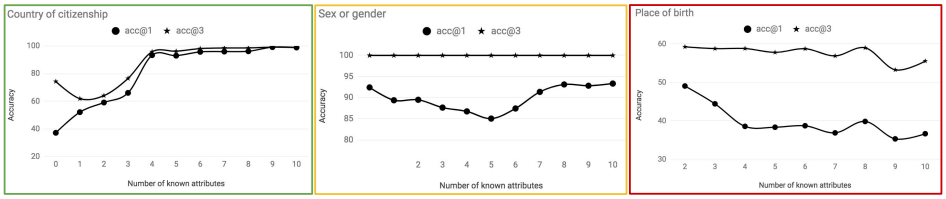


Figure 32: Dependency of the accuracy of profiling politicians on the number of known facets: a positive correlation for *country of citizenship* (left Figure), no correlation for *sex or gender* (center), and a slightly negative one for *place of birth* (right).

Table 48 provides the results of our methods and baselines on the three smaller datasets that contain embeddings (the full datasets yielded similar results for MFV, NB, and AE). We observe that AE and EMB outperform the baselines on almost all cases, as hypothesized in P1. As expected (hypothesis P3), we see lower (or no) improvement over the baselines for cases with low entropy (e.g., *sex or gender* and *lifespan range*) compared to attributes with high entropy (e.g., *award received*). We also note that the accuracy of profiling per facet correlates inversely with its vocabulary size v_i (hypothesis P4).

The superiority of the neural models over the baselines means that capturing complex inter-facet dependencies improves the profiling ability of machines. Moreover, although the two neural methods perform comparably on average, there are differences between their accuracy on individual facets (e.g., compare *award received* and *native language* on any dataset). To gain further insight, we analyze the individual predictions of our models on an arbitrarily chosen instance from our dataset, namely, the Ohio politician Brad Wenstrup.[12]. Brad is a male American citizen, member of the republican party, born in Cincinnati, Ohio in the 20th century, educated at the University of Cincinnati, and has held the position of a United States representative. Both neural systems correctly predict (as a first choice) Brad's country of citizenship, position held, work location, and century. The AE system is able to predict the political party and his education better than the EMB one; whereas the EMB model is superior over AE when it comes to Brad's gender. In addition, EMB ranks the correct place of birth third in terms of probability, while AE's top 3 predictions for this attribute are incorrect, i.e., these are places in the neighboring state of Michigan. These differences might be due to the main architectural difference between these two methods: EMB's input embedding contains much more information (both signal and noise) than what is captured by the 14 facets in the AE.

How does a profile improve (or not) with increasing input? To investigate our hypothesis P5, we analyze both top-1 and top-3 accuracies of AE for predicting a facet value against the number of other known facets provided at test time. Figure 32 shows examples of all three possible correlations (positive, negative, and none) for politicians. These findings are in line with conclusions from social psychology (cf. *Profiles*): knowing more facets of an instance might trigger a shift

12 https://www.wikidata.org/wiki/Q892413

of the original profile, and it might be correct or incorrect, as defined in (Stone et al., 1957). Generally, we expect that attributes with large v_i, like *place of birth*, will suffer as input exemplars become more specified and granularity becomes tighter, while facets with small v_i would benefit from additional input. Figure 32 follows that reasoning, except for *sex or gender*, whose behavior is additionally influenced by low entropy (0.25) and strong frequency bias to the *male* class.

6.7.2 *Comparison against human expectations*

In order to collaborate with humans, or understand human language and behavior, both humans and AI machines are required to fill knowledge gaps with assumed world expectations (see section 6.1). Given that in most AI applications information is created for humans, a profiler has to be able to mimic human expectations. We thus compare our neural profiles to profiles generated by crowd workers.

6.7.2.1 *Data*

We evaluate on 10 well-understood facets describing American citizens. For each facet, we generated a list of 10 most frequent values among American citizens in Wikidata, and postprocessed them to improve their comprehensibility. We collected 15 judgments for 305 incomplete profiles with the Figure Eight crowdsourcing platform. The workers were instructed to choose 'None of the above' when no choice seemed appropriate, and 'I cannot decide' when all values seemed equally intuitive. We picked reliable, US-based workers, and ensured US minimum wage ($7.25) payment.

Given that there is no 'correct' answer for a profile and our annotators' guesses are influenced by their subjective experiences, it is reasonable that they have a different intuition in some cases. Hence, the relatively low mean Krippendorff (1980) alpha agreement per property (0.203) is not entirely surprising. We note that the agreement on the high-entropy attributes is typically lower, but tends to increase as more facets were provided. Overall, the annotators chose a concrete value rather than being indecisive ('I cannot decide') for the low-entropy more often than the high-entropy facets. When more properties were provided, the frequency of indecisiveness on the high-entropy facets declined.

6.7.2.2 *Results*

When evaluating, 'None of the above' was equalized to any value outside of the most frequent 10, and 'I cannot decide' to a $(1/N)$-th vote for each of the values. The human judgments per profile were combined in a single distribution, and then compared to the system distribution by using Jansen-Shannon divergence (*JS-divergence*).[13] We evaluate the profiles generated by our AE and the

13 We considered the following metrics: JS-divergence, JS-distance, KL-divergence, KL-divergence-avg, KL-divergence-max, and cosine distance (Mohammad and Hirst, 2012). The agreement was very high, the lowest Spearman correlation being 0.894.

baselines; EMB could not be tested on this data since most inputs do not have a corresponding Wikipedia page and pre-trained embeddings.

The divergence between our AE system and the human judgments was mostly lower than that of the baselines (Table 49), which supports our expectation in the hypothesis P2. The divergences for any system have a strong correlation with (normalized) entropy, confirming our previous observation that high-entropy attributes pose a greater challenge (hypothesis P3). We also computed precision, recall, and F1-score between the classes suggested by our system and by the annotators, and observed that it correlates inversely with the entropy in the data (H_i), as well as the entropy of the human judgments (J_i).

The results show that our AE can capture human-like expectations better than the two baselines, and that mimicking human profiling is more difficult when the entropy is higher. While parameter tuning and algorithmic inventions might improve the profiling accuracy further, it is improbable that profiles learned on factual data would ever equal human performance. Some human expectations are rather 'non-factual', i.e., they are culturally projected and do not correspond to episodic facts. Future work should seek novel solutions for this challenge.

6.8 DISCUSSION AND LIMITATIONS

6.8.1 *Summary of the results*

Identity clustering is nearly ideal assuming perfect attribute extraction, indicating that to a large extent the information available in text is sufficient to establish identity, as long as the ambiguity is low (hypothesis C1). As expected, the clustering performance declines when the attribute extraction is imperfect (C2).

Given that the clustering based on the gold properties is relatively high, there is little room for improvement by profiling in this case. It is thus no surprise that the profiler has no visible effect when combined with gold properties (C3). Profiling is able to fill certain knowledge gaps when combined with automatic extraction of properties (C4). Concretely, profiling consistently has a positive impact when using exact reasoning over property values. Profiling is also beneficial when applied together with the no-clash reasoner on the full datasets (FD1-FD3), whereas it performs slightly worse than the no-clash reasoner on the partial datasets (PD1-PD4). Overall, we observe that the profiler is fairly robust to different hyperparameter values and degrees of data ambiguity.

When analyzing these results, we need to be mindful about the low performance of our automatic attribute extractors, especially in terms of recall. Considering that the effectiveness of our profiler is directly conditioned on the completeness and accuracy of its input, the generated profiles might not be useful when the input information is scarce or inaccurate. Incorrect or incomplete inputs lead to more generic or even wrong profiles, and the decisions on identity based on these profiles are likely to be consequently wrong as well.

The datasets with larger ambiguity pose a greater challenge for all approaches (C5). However, we did not see a clear relation between the usefulness of our

profiler and the data ambiguity (C6). A possible explanation for this finding lies in the low number of properties considered, as well as in their usefulness to discriminate long-tail entities in the test domain. Future work should investigate whether these findings generalize for an extended set of properties.

The intrinsic evaluation of our profiling methods demonstrated their ability to largely mimic the instantial data in Wikidata (P1), as well as human judgments (P2). Notably, their performance per attribute fluctuates dramatically, but this variance can be easily predicted by the factors of entropy, value size, and other known attributes (hypotheses P3, P4, and P5). It remains a future task to investigate the impact of this accuracy variance on the clustering performance of the profiler.

6.8.2 *Harmonizing knowledge between text and knowledge bases*

A large challenge during these experiments lay in harmonizing the knowledge found in the text documents with the one found in the chosen knowledge base, Wikidata. We discuss three aspects of this challenge here, as these were notable limitations to our experimental setup and we expect them to be relevant for future research of similar nature.

1. **Discrepancy of properties** There is little overlap between the attributes that are found in background knowledge and those found in text. For example, descriptions of the kind of area in which an entity lives (e.g., whether a city part is dangerous or safe), are prominent in text and would be very useful to establish identity of entities; unfortunately, this kind of information is not present in Wikidata. On the other hand, one's height or political affiliation is often stored in Wikidata, but it is not mentioned in text.

2. **Discrepancy of property values** The values for an attribute that are found in text are often dissimilar with those in the background knowledge base. For instance, most causes of death in Wikidata are natural and do not correspond to those in the gun violence domain, where most people died of a gunshot wound. In addition, the number of values for certain properties, such as birthplace, in Wikidata is quite large. For each property, we opted to choose between several (typically between two and ten) attribute values on a coarser granularity level. It was not trivial to map the original property values in Wikidata to coarser classes, as the connection in the structured knowledge was not consistently available. For instance, the tens of thousands of places of birth could not be automatically mapped to states of birth, because the connection between these two is not consistently stored in Wikidata.

3. **Discrepancy of world expectations** Even after the attributes and some of their values have been harmonized, it can be the case that the expectations learned from a knowledge base are not fitting those within a local crime domain. As an example, let us consider the attribute *education level*. People

in Wikidata are typically highly educated, which is probably not representative for the typical education level of the participants in the gun violence incidents. Another example is the property *occupation*: typical professions in Wikidata, such as politicians or actors, are unlikely to be encountered in gun violence incident descriptions.

These three discrepancies are largely due to the different world that is modeled between these two proxies: while our documents describe local crime incidents, the background knowledge in Wikidata models global events and well-known people. Learning models about the 'head' entities in Wikidata and applying these to the 'long-tail' entities in the gun violence domain required a substantial amount of engineering and has been achieved at the expense of information loss, which largely decreased the discriminative power of our profiling machines. Future work should investigate directions to learn expectations that would resemble the target domain closer, e.g., using gun violence domain data to both train the profiles and apply them.

6.8.3 *Limitations of profiling by NNs*

Our experiments show the natural power of neural networks to generalize over knowledge and generate profiles from data independent of schema availability. Techniques like dropout and oversampling further boost their ability to deal with missing or underrepresented values. Ideally these profiling machines can be included in an online active representation system to create profiles on the fly, while their modularity allows easy retraining in the background when needed.

Still, it is essential to look critically beyond the accuracy numbers, and identify the strengths and weaknesses of the proposed profiling methods. Limitations include: 1. **continuous values**, such as numbers (e.g., age) or dates (e.g., birth date), need to be categorized before being used in an AE;[14] 2. AE cannot natively handle **multiple values** (e.g., people with dual nationality). We currently pick a single value from a set based on frequency; 3. as noted, we applied dropout and oversampling mechanisms to reinforce **sparse attributes**, but these remain problematic; 4. it remains unclear **which aspects of the knowledge** are captured by our neural methods, especially by the EMB model whose embeddings abstract over the bits of knowledge. More insight is required to explain some differences we observed on individual facets.

6.9 CONCLUSIONS AND FUTURE WORK

In this chapter, we investigated the role of knowledge when establishing identity of NIL (E3) entities mentioned in text. Having no representation in DBpedia or other common knowledge bases in the LOD cloud, the NIL entities are different from head and tail entities for the following reasons: 1. they contain no frequency/popularity priors 2. the knowledge on them is scarce and non-redundant

14 We obtained *lifespan* and *century of birth* from birth and death dates.

3. their ambiguity is unknown, but potentially extremely large. Despite these unique properties, the NIL entities have mostly been out-of-focus in existing research that establishes identity of entities from text.

Considering the three properties of the NIL entities, establishing their identity is a knowledge-intensive task, requiring the explicitly stated information in text ('local context') to be combined with implicit expectations based on background knowledge. In this work, we formalized the local context through a set of property-value facts, based on intratextual knowledge extracted from text. The global expectations, which we called *profiles*, were represented with probability distributions, learned by computational models over existing extratextual knowledge.

Inspired by the functions of profiles in human cognition and the needs for processing long-tail identity in NLP, we posed the question: how can we naturally fill knowledge gaps in communication with assumed expectations. Existing research in KBC cannot be used for this purpose, since they focus on predicting concrete missing facts (such as exact age or location) for known entities. Therefore, we proposed profiling, a more flexible and robust variant of KBC that generates distributions over ranges of values. We described two profiling machines based on state-of-the-art neural network techniques.

We put forward 6 hypotheses about the role of background knowledge in the task of establishing identity of NIL entities in text. Imperfect (automatic) attribute extraction led to lower performance in comparison to gold extraction. Profiling was able to fill certain gaps of the automatic extractors: profiling consistently helped the exact reasoner, whereas it performed comparably to the no-clash baseline. As expected, higher ambiguity made the task much harder. However, the usefulness of profiling had no clear relation to the degree of data ambiguity. The results obtained so far might have been affected by the low quality (precision and recall) of the automatic extractors, as profiles built on top of incomplete or wrong input might amplify this input and will likely harm the identity clustering decisions. It remains to be seen whether the current findings persist when more properties or better automatic extractors are employed.

The profiles are explicit representations, thus providing us with transparency and explanation on the identity decisions. To further understand the behavior of the profilers, we performed two intrinsic experiments of the profiling methods we propose in this chapter, thus testing another 5 hypotheses. Namely, we evaluated the profiling machines against instantial data in Wikidata, as well as against human judgments collected in a crowdsourcing task. Both experiments showed that the profiling methods get much closer to the typical or the correct value than the underlying baselines based on frequency or the Naive Bayes method. As hypothesized, the prediction accuracy per attribute varies greatly, but this can be largely explained through the notions of entropy, value size, and (number of) known attributes.

A large challenge during these experiments lay in harmonizing the knowledge found in the text documents with the kind of information found in the background knowledge base. Firstly, there is little overlap between the attributes

that are found in background knowledge and those found in text. Secondly, the values for those attributes that are found in text are often dissimilar with those in the background knowledge base. This is largely due to the different world that is modeled between these two proxies: while our documents describe local crime incidents, the background knowledge base models global events and well-known people. In addition, even after the attributes and some of its values are harmonized, it can be the case that the expectations learned from a knowledge base are not fitting those in a local crime domain.

This knowledge-intensive approach provided a valuable link between the local context, which is currently based on intratextual knowledge, and the profiles, learned on top of extratextual knowledge. Subsequent work should investigate the added value of integrating knowledge from the circumtextual and intertextual categories, when establishing identity of long-tail entities.

Answer to RQ5 This chapter aimed to provide evidence to the fifth research question in this thesis: *What is the added value of background knowledge models when establishing the identity of NIL (E3) entities?* For this purpose, we proposed to establish the identity between two local representations of an entity by combining explicit and implicit knowledge about their properties. Given that the available information in text could be insufficient to establish identity, we enhanced it with profiles: background knowledge models that aim to capture implicit expectations left out in text, thus normalizing the comparison and accounting for the knowledge sparsity of most attributes. We systematically investigated the role of different pieces of explicit knowledge, reasoning methods, as well as the intrinsic behavior of the profiling methods. We started this investigation by matching the local context to pre-built profiles based on the extratextual knowledge in Wikidata. This creates an interesting opening for future research to investigate the other types of knowledge, i.e., intertextual and circumtextual knowledge.

This research has thus advanced the understanding of the under-addressed challenge of establishing identity of the NIL entities found in text, as well as of the role of implicit extratextual knowledge in resolving this task. The results obtained so far showed certain impact of using profiling on top of the attributes extracted automatically from text, whereas the relation of this impact to data ambiguity is unclear at this point. They also revealed potential pitfalls that need to be considered in future work, namely: 1. profiles built on top of incomplete or wrong information extracted from text can be misleading and counterproductive for the task of establishing long-tail identity 2. there is a notable discrepancy between the properties and their values found in text in comparison to those that can be learned from Wikidata. Overall, we argue that the general assumption is correct and worth pursuing further, whereas certain aspects of the experimental design should be revisited in future research.

From this perspective, it seems that further work should continue investigating how to best incorporate profiling components to fill the gaps in human communication. With further understanding and engineering of these phenomena, we expect that such profiling machines would be able to natively address (at least) three standing problem areas of modern-day NLP: 1. scarcity of episodic

knowledge, prominent both in knowledge bases and in communication; 2. unresolved ambiguity in communication, when the available knowledge is not necessarily scarce, yet prior expectations could lead to more reliable disambiguation; 3. anomaly detection, when a seemingly reliable machine interpretation is counter-intuitive and anomalous with respect to our expectations.

	PERSON				*POLITICIAN*				*ACTOR*			
attribute	$\mathbf{n_{ex}}$	$\mathbf{v_i}$	$\mathbf{H_i}$	$\mathbf{H_i\prime}$	$\mathbf{n_{ex}}$	$\mathbf{v_i}$	$\mathbf{H_i}$	$\mathbf{H_i\prime}$	$\mathbf{n_{ex}}$	$\mathbf{v_i}$	$\mathbf{H_i}$	$\mathbf{H_i\prime}$
educated at	273,096	3,000	9.28	0.80	22,461	3,000	9.73	0.84	5,047	883	7.56	0.77
sex or gender	2,403,980	11	0.64	0.18	168,758	5	0.50	0.25	75,980	5	1.00	0.50
citizenship	1,546,757	995	5.28	0.53	152,131	335	5.07	0.61	57,570	187	5.12	0.68
native language	41,760	141	1.70	0.24	16,818	33	1.08	0.21	4,273	29	0.41	0.08
position held	177,302	3,000	7.44	0.64	101,766	1,701	7.08	0.66	244	25	0.96	0.21
award received	154,275	3,000	7.97	0.69	10,588	546	6.82	0.75	2,880	297	6.60	0.80
religion	32,311	341	3.24	0.38	2,414	127	3.99	0.58	164	24	2.47	0.56
political party	158,105	3,000	7.28	0.63	82,617	2,456	7.26	0.64	232	53	3.23	0.58
work location	68,602	1,989	6.25	0.57	30,320	272	5.07	0.63	116	41	3.99	0.74
place of death	350,720	3,000	7.93	0.68	29,071	3,000	8.39	0.73	9,377	2,169	8.33	0.75
place of birth	927,089	3,000	7.64	0.66	59,627	3,000	7.27	0.63	39,694	3,000	8.55	0.74
cause of death	21,926	499	5.35	0.60	1,408	115	4.75	0.69	1,039	82	4.22	0.66
lifespan range	922,634	55	1.89	0.33	79,346	39	1.68	0.32	19,055	11	1.77	0.49
century of birth	1,975,197	43	1.36	0.25	140,087	22	1.48	0.33	61,506	11	0.56	0.16

Table 47: Numbers of examples (n_{ex}), categories (v_i), and entropy (H_i and $H_i\prime$) per facet of People in our training data. We limit v_i to 3,000 to restrict the complexity of the value space, but also to mimic the simplification aspect of cognitive profiling.

	PERSON				*POLITICIAN*				*ACTOR*			
attribute	**MFV**	**NB**	**AE**	**EMB**	**MFV**	**NB**	**AE**	**EMB**	**MFV**	**NB**	**AE**	**EMB**
educated at	4.41	9.22	13.20	**22.45**	2.57	6.88	**13.14**	9.47	11.32	15.09	3.77	**46.43**
sex or gender	82.61	81.76	82.37	**95.83**	85.15	84.10	83.23	**94.79**	49.71	57.97	55.20	**89.06**
citizenship	29.10	57.36	66.49	**78.49**	18.27	46.75	72.94	**77.96**	17.99	39.94	60.77	**65.05**
native language	44.70	69.44	**87.63**	33.33	46.67	88.89	**93.33**	83.33	**95.00**	**95.00**	**95.00**	91.67
position held	8.44	32.92	**45.66**	21.43	15.47	28.93	**45.03**	41.18	50.00	50.00	50.00	**100.0**
award received	4.98	15.95	21.56	**37.50**	3.85	10.58	18.27	**26.09**	14.29	14.29	23.81	**42.86**
religion	27.52	40.83	45.48	**71.43**	27.08	42.71	52.08	**56.52**	40.00	40.00	60.00	**66.67**
political party	13.18	29.67	42.08	**47.06**	9.41	22.78	34.28	**37.59**	**50.00**	**50.00**	**50.00**	0.0
work location	22.47	57.18	**64.49**	60.00	22.22	69.90	**83.09**	75.00	0.00	0.00	0.00	0.00
place of death	4.09	25.09	28.20	**36.84**	2.81	8.03	17.27	**25.81**	9.78	17.58	18.48	**33.93**
place of birth	2.85	33.01	32.07	**49.04**	1.88	**54.62**	23.59	52.21	5.31	11.28	16.87	**36.21**
cause of death	23.80	24.13	**24.24**	15.38	32.76	37.93	24.14	**71.43**	33.33	33.33	20.00	**45.00**
lifespan range	41.76	**43.56**	41.69	42.03	41.30	40.68	38.51	**48.75**	36.73	39.17	**45.92**	43.33
century of birth	82.04	85.45	84.94	**89.53**	76.13	80.13	83.14	**85.79**	**93.62**	93.60	89.56	92.67

Table 48: Top-1 accuracies for the both neural methods and the two baselines on the smaller datasets. For each dataset-facet pair, we emphasize the best result. Our neural methods, especially EMB, outperform the baselines. Entropy and vocabulary sizes can partially explain deltas in accuracies on individual facets.

Table 49: Human evaluation results per attribute: number of values (v_i), entropy (H_i), normalized entropy ($H_i\prime$), mean judgments entropy (J_i), divergences of: MFV, NB, and AE.

attribute	v_i	H_i	$H_i\prime$	J_i	**MFV**	**NB**	**AE**
cent. of birth	5	0.40	0.92	10^{-8}	0.13	**0.12**	**0.12**
religion	4	0.63	1.26	10^{-10}	**0.05**	0.09	0.06
sex or gender	2	0.70	0.70	10^{-14}	0.04	**0.02**	**0.02**
place of death	8	0.80	2.40	0.05	0.51	0.20	**0.16**
lifespan range	10	0.81	2.68	0.02	0.29	**0.09**	**0.09**
place of birth	8	0.83	2.48	0.01	0.39	0.26	**0.24**
work location	10	0.84	2.80	0.03	0.49	**0.28**	0.30
occupation	9	0.92	2.90	0.06	0.37	0.36	**0.32**
educated at	9	0.92	2.91	0.06	0.39	0.25	**0.23**
political party	2	1.00	1.00	0.02	0.17	**0.06**	**0.06**

7 CONCLUSION

7.1 SUMMARIZING OUR RESULTS

How can the performance of NLP techniques to establish identity of long-tail cases be improved through the use of background knowledge? In this thesis I approached this research question by focusing on multiple aspects in parallel:

1. **RQ1 (Description and observation):** *How can the tail entities be distinguished from head entities?*
2. **RQ2 (Analysis of the evaluation bias):** *Are the current evaluation datasets and metrics representative for the long-tail cases?*
3. **RQ3 (Improvement of the evaluation bias):** *How can we improve the evaluation on the long-tail cases?*
4. **RQ4 (Access to knowledge):** *How can the knowledge on long-tail entities be accessed and enriched beyond DBpedia?*
5. **RQ5 (Role of knowledge):** *What is the added value of background knowledge models when establishing the identity of NIL entities?*

I will next review the challenges encountered with respect to each of these aspects, our approach to address them, and the main findings.

7.1.1 *Describing and observing the head and the tail of Entity Linking*

Problem Every entity belongs to one of the following three groups: head (E1) entities, tail (E2) entities, or NIL (E3) entities, based on its lexical usage, frequency in communication, and popularity. However, whereas it is easy to distinguish NIL entities, it is much more challenging to decide between the head and the tail cases, since the head-tail distinction has never been described so far. Without a concrete description of the head and the tail cases, it is not possible to quantify their potential impact on system performance, and consequently, it is impossible to address the difficult cases of the entity linking task.

Approach My aim was to provide a description of the head and the tail in entity linking, as well as its impact on system performance. For that purpose, I first defined the head-tail properties of entity linking along which the head and the tail cases can be distinguished. I proposed 16 hypotheses that describe how these properties interact in EL datasets and how they relate to system performance. I tested these hypotheses on three contemporary EL systems and five corpora.

Conclusion This is the first systematic study of the relation between surface forms in EL corpora and instances in DBpedia, based on a set of hypotheses on what long tail phenomena are. The analyses of EL corpora with respect to these properties revealed certain correlations that followed the posed hypotheses. Most importantly, the head-tail phenomena and their interaction consistently predict system performance. Namely, there is a positive dependency of system performance on frequency and popularity of instances, and a negative one with ambiguity of surface forms. Essentially, this confirms the intuition that system performance is largely based on head cases, and declines strongly on the tail. To support the inclusion of the tail in future designs of EL systems and datasets, I provided a set of recommended actions to be considered when creating a dataset, evaluating a system, or developing a system in the future.

7.1.2 *Analyzing the evaluation bias on the long tail*

Problem We observed that system performance declines as we move from the head towards the tail of the entity distribution. At the same time, the overall system performance measured on existing datasets is fairly high, and comparable to their performance on the head cases. This signals that existing datasets are expected to mostly evaluate on the head of the entity distribution. Considering that the design and annotation of a new dataset by itself is an extremely laborious process, it is understandable that certain unmotivated biases exist in these datasets. At the same time, it is crucial that each NLP task is tested on representative data, or at least, that when evaluating we are aware of the specific properties of the dataset and what it actually assesses.

Approach Hence, I set a goal to understand the representativeness of existing datasets used to evaluate systems on disambiguation and reference tasks. For that purpose, I provided a world model that defines these tasks through a mapping between surface forms found in corpora and their meaning (concept or instance) as found in a resource. Based on this model, I proposed a set of metrics that quantify the form-meaning relations and applied these metrics on existing datasets from five tasks: Entity Linking, Entity Coreference, Word Sense Disambiguation, Semantic Role Labeling/Verb Disambiguation, and Event Coreference.

Conclusion The analysis showed that the commonly used datasets to evaluate these tasks suffer from low ambiguity, low variance, high dominance, and limited temporal spread. Whereas this analysis did not yet reveal how to create datasets which fill the gap, it did make us aware of measurable semantic areas to which datasets, and consequently systems, overfit. As such, it can be seen as a specification towards complementing the existing test cases with novel datasets that evaluate different aspects of the spectrum. Moreover, the metrics proposed in this work could be used as a driving mechanism to create new metric- and time-aware datasets in a systematic manner - or at least, to assess and revise their representativeness upon creation.

7.1.3 *Improving the evaluation on the long tail*

Problem Since the tail is underrepresented in current datasets, the high results that are achieved by state-of-the-art systems can be mostly attributed to head cases. This means that our systems can interpret generally known/popular entities and events. At the same time, while we do not have a good understanding about how these systems perform on the tail, the limited evidence available so far shows that understanding local/contextual entities and events is beyond their capabilities. It is important to understand and overcome these challenges, especially because most of the world consists of tail cases, with their unique particularities.

Approach Hence, I focused on creating the first task that deliberately evaluates the tail of disambiguation and reference tasks. This was done by designing a higher-level task of Question Answering, in such a way that requires systems: 1. to combine the knowledge from different NLP disambiguation and reference tasks 2. to do so on local events and unpopular entities, both with no representation in common knowledge bases 3. to operate within challenging linguistic settings of high ambiguity and variance. During the creation of this task, a novel method called "data-to-text" was invented. This method relies on existing repositories having a link between structured data summarizing an event/incident, and text documents providing evidence for that data. Data-to-text allows one to create datasets in a semi-automatic way, with minimal post-validation.

Conclusion Using the data-to-text method, I generated a lot of data covering the long tail with minimal post-validation. This data was filtered in order to gain a lot of confusability and to challenge systems to perform deep interpretation on long tail data. I carried out this work as organizer of a task at the SemEval-2018 competition, entitled "Counting Events and Participants within Highly Ambiguous Data covering a very long tail".[1] Four systems participated in it, reaching to at most around 30% accuracy on answering the task questions, showing that dealing with high ambiguity and not being able to rely on frequency biases, poses a great challenge for the current NLP technology. As the challenges relating to interpreting long-tail cases become more recognized in our community, it is to be expected that this task will attract interest in the following years. To support this, an easy evaluation for out-of-competition systems was enabled. Finally, the "data-to-text" method used to create this task, could be applied again to create much more long tail data, covering other domains and spatio-temporal contexts.

To create this task, we deliberately focused on long-tail data with high confusability, without preserving the natural distribution of the texts. We see this aspect of artificiality as a necessary trade-off in order to achieve our goal: to create a task that deliberately targets the evaluation bias on the tail, resulting in large confusability and requiring an extremely low annotation effort.

1 https://competitions.codalab.org/competitions/17285

7.1.4 *Enabling access to knowledge on the long-tail entities*

Problem We observe a similar bias in knowledge sources as we do in evaluation datasets. Tail (E2) and NIL (E3) entities are characterized by scarce to no accessible knowledge at all, in the commonly used, Wikipedia-based knowledge bases, such as DBpedia. The lack of background knowledge to reason over prevents successful linking of tail entities, because in such cases it is extremely challenging to compete with head interpretations that share the same form. At the same time, the Linked Data community provides us with hundreds of thousands of data documents, covering billions of facts about various entities, both contextually relevant and generally known ones. Unfortunately, the majority of the Linked Data instances and statements have not found their way in the entity linking task so far. This can be attributed to multiple reasons, the primary obstacle being *accessibility* - today it is not straightforward to access these statements and use them in an EL system.

Approach It is hence crucial to improve the access to this knowledge found in the Linked Data cloud. To do so, I built on top of an existing data cleaning architecture and collection of RDF statements, the LOD Laundromat. The LOD Laundromat stores tens of billions of RDF statements in a central repository, on disk, and allows them to be queried via the LDF triple pattern matching mechanism. Since this architecture did not include a text search feature, I teamed up with its authors and enriched the Laundromat with a centralized text index and search tool, named LOTUS. LOTUS indexes over 4 billion RDF literals, through which one can access billions of URIs and hundreds of thousands of documents. LOTUS allows flexible retrieval, by offering the user to choose between 32 combinations of matching and ranking algorithms. It relies on ElasticSearch and LOD Laundromat, and has been distributed over a number of simple servers to ensure scalability.

Conclusion With this approach, I was able to fill the gap and provide the LOD community with the largest centralized text index and access point to the LOD Laundromat data collection. Moreover, this allows EL systems to use the knowledge found among the billions of statements of the LOD Laundromat collection. This represents a first, but crucial step, towards building EL systems that can leverage the richness and diversity of the knowledge found on the Web. Initial experiments on three use cases show that LOTUS holds a potential to increase the recall of the entity linking systems. Namely, we observed that the proportion of DBpedia resources in the top 100 results is between 22% (for querying for names of journals) and 85% (when querying for local monuments). A subset of the queries only returned candidates that are found outside of DBpedia.

A general increase of the recall is desired for the tail cases, but by doing so, the precision of the entity linking decreases, in particular on the head cases. Future work should investigate how to balance this newly obtained recall with precision, as well as how to evaluate a web-of-data-wide entity linking system.

7.1.5 *The role of knowledge in establishing identity of long-tail entities*

Problem For E2 entities we can find instantial representation and associated knowledge beyond DBpedia in the Linked Data cloud. For many others (NIL/E3 entities), however, we have no accessible existing representation at all. The lack of frequency priors among these instances (I), the scarcity and non-redundancy of knowledge (II), and the unknown, but potentially extreme, ambiguity (III) make the NIL entities substantially different than the head entities. How to establish and represent the identity of these entities? What kind of knowledge can be applied to establish identity of NIL entities? This is currently an under-addressed research area.

Approach An intelligent interpretation of long-tail cases requires a revision and extension of contextual knowledge, as well as an approach to dynamically acquire such knowledge for each long-tail case. I investigated the role of explicit knowledge extracted from text and implicit models learned from extratextual background knowledge, for establishing identity of long-tail, NIL entities mentioned in text. Perfect extraction of information from text could be sufficient for this purpose, provided that the right information is available. Given that the information is not consistent across documents and that automatic extractors are far from perfect, I attempted to use background knowledge models ("profiling machines") in order to complete the partial profiles extracted from text. Profiling is inspired by human capabilities of stereotyping, and belongs to the Knowledge Base Completion family of tasks. Two neural architectures were developed for this purpose. I investigated six hypotheses regarding the usefulness of knowledge when establishing identity of NIL entities, and five additional hypotheses on the intrinsic behavior of the profiling machines.

Conclusion The profiling machines were able to mimic instance data and human judgments to a large extent. The evaluation of these machines on the task of establishing long-tail identity in text showed promising results when applied on top of automatically extracted information from text. I observed no clear patterns between the effectiveness of these profilers and the data ambiguity. My experiments also revealed potential pitfalls that need to be considered in future work, namely: 1. profiles built on top of incomplete or wrong information extracted from text can be misleading and counterproductive for the task of establishing long-tail identity 2. there is a notable discrepancy between the properties and their values found in text in comparison to those that can be learned from Wikidata. Overall, I argue that the general assumption is correct and worth pursuing further, whereas certain aspects of the experimental design should be revisited in future research.

While profiles generated from facts help extrinsic tasks and improve over simple baselines, they still lack the knowledge belonging to the circumtextual and the intertextual category. In addition, generating profiles that exactly correspond to human expectations would require them to have access to non-factual human beliefs as well. These challenges should be addressed in future work.

7.2 LESSONS LEARNED

This thesis systematically investigated how to establish the identity of long-tail entities from a perspective of a computational NLP system. I made an effort to describe these entities, to analyze their presence in current evaluations, to improve their evaluation bias, to increase the access to knowledge about them, and to establish their identity when a representation is not readily available at all. I believe that the lessons learned in this thesis can be applied to a number of other disambiguation and reference tasks in NLP, given that the head-tail distinction and challenges are present in essentially most (if not all) NLP tasks we deal with.

To help future related research, I next discuss some key lessons learned throughout the course of my PhD research, expressed as 6 observations and 6 recommendations.

7.2.1 *Observations*

The first two observations regard the system performance on long-tail cases.

1. *Performance scores mean very little without further interpretation* Today, NLP revolves around the goal of obtaining high scores and 'beating the state-of-the-art'. Without a doubt, obtaining the best result so far has its own benefits, and provides a claim that a certain system is superior over lower-performing ones. This thesis shows that the significance of these scores on their own is not clear without further understanding of the underlying behavior, given the small size and the lack of representativeness of current evaluation datasets. In order to have any meaningful interpretation of the obtained F1-scores, they must be paired with certain detail on the properties of the evaluation data, typical cases on which the system performs well, and typical errors that the system makes. Only then we would be starting to get an idea about the strengths and weaknesses of these systems, and about the limits of their applicability.

2. *The tail performance is not derivable from the head performance* The fact that the performance of systems on the head and the tail is largely different, implies that there is not only a quantitative, but also a qualitative difference between these two parts of the entire entity distribution. For this purpose, it is essential to perform evaluation on the head and the tail separately, as well as together, and analyze the findings in detail. Such a dedicated evaluation would potentially inspire a stronger focus on the long-tail cases in future research.

I next present two observations concerning the evaluation of long-tail identity.

3. *Establishing entity identity is far from solved today* One of the main points of this thesis is that the task of establishing identity of entities is far larger

than the currently dominant way of entity interpretation from text, namely by Entity Linking. Whereas systems can perform well on the few head cases, the performance on the far more numerous tail entities is much lower. Furthermore, systems can typically not deal with entities that have no accessible representation at all on the web.

4. *Under-addressed problems require much more than only developing a good processor* Each novel challenge requires much more than only developing a system approach that attempts to provide a solution for it. System development is/should ideally be preceded by a formal definition/description of the task, and good understanding of the state-of-the-art evaluation and methods with respect to that specific aspect. For that purpose, this thesis combines five aspects, ranging from a problem description to development of a knowledge-based system, in an attempt to advance the state-of-the-art on the tail cases of establishing entity identity. I took a breadth-first approach, because I believe that these aspects should progress in parallel, and they are all important for the advancement of our understanding of the long-tail phenomena when establishing identity of entities. Notably, addressing various, complementary dimensions of such novel challenges should be seen as a joint, community effort.

The last two observations are about the knowledge used in current systems.

5. *Systems miss a lot of knowledge* The classification of framing and interpretation mechanisms used by humans to interpret text is an eye-opener, demonstrating that systems entirely miss certain types of knowledge (e.g., circumtextual), whereas other types of knowledge are only integrated to a limited extent (e.g., extratextual knowledge). As demonstrated to some extent in this thesis, we can systematically trace the impact of this missing knowledge on individual cases.

6. *Accessing missing knowledge is non-trivial* There might be a very practical explanation on why certain knowledge types are generally well-represented, whereas others are constantly absent. Namely, accessing certain types of knowledge might depend on significant preparatory work. Two examples of this can be found in this thesis. Firstly, in order to start considering Entity Linking over the Linked Data cloud, we first need a textual index that allows us to access this data in the first place - this is the motivation for the creation of LOTUS in chapter 5. Secondly, in order for us to be able to use a module that generates stereotypical expectations about people, I had to first investigate, design, and implement this as a separate profiling component.

7.2.2 *Recommendations*

The first set of recommendations discusses the distributional tail as an objective for systems.

1. *Optimize for the tail!* Given that current EL systems struggle to interpret tail cases, it is intuitive that the tail should receive more attention in future designs of systems. To my knowledge, no existing system has deliberately focused on interpreting tail cases.

2. *Enhance systems with missing background knowledge* Our investigation of the EL systems through the knowledge used demonstrated that much knowledge available to humans is not available or leveraged by state-of-the-art systems. Future designs of systems should take this into account. Integrating circumtextual and intertextual knowledge is desired as these two knowledge types are utterly absent in current systems. Similarly, extratextual, 'world' knowledge can be further exploited, e.g., directly to perform reasoning on anomalous interpretations, but also indirectly to induce procedural "script" knowledge or stereotypical profiles (like our attempt in chapter 6). Intuitively, using more extensive background knowledge is most relevant for long-tail cases, that lack instance-level knowledge and cannot be resolved by relying on (frequency) priors.

3. *Combine deep/machine learning with linguistics and pragmatics* Theories and propositions stemming from linguistics and pragmatics might not have found their place entirely in state-of-the-art reference systems yet, but they contain a lot of value. This is somewhat a side-effect of the vast advent of deep neural networks in the recent years. While neural networks have advanced the state-of-the-art performance of systems, they are not almighty and there is space for integrating them with findings from other disciplines, such as linguistics and pragmatics. After all, the human brain is not (solely) a statistical machine, as it does not expose the same frequency bias as our machine learning algorithms. Attending to the long-tail cases is an attempt to capture this amazing capability of the human brain to be less primed by statistical priors, and reconcile them optimally with the right knowledge in a given context.

Next, I present recommendations that can be applied to improve the evaluation on the distributional tail.

4. *Use macro F1-score helps to alleviate the frequency bias of the head* As concluded in (Ilievski et al., 2018), one way to decrease the effect of frequency in existing datasets is to apply a macro F1-score instead of micro F1-score when measuring system performance. The rationale behind this proposal is that macro F1-score evaluates each (form or instance) class once irrespective of the number of its occurrences.

5. *Make use of informative baselines* We need to be aware of the extent to which relevant baselines, especially simple ones like a most-frequent value strategy, can perform on a certain dataset. Certainly, system accuracy of 75% is far more impressive when the baseline score is 30% compared to 70%.

Not considering this aspect makes us vulnerable to situations where a system combined a large set of features, but the performance of using only PageRank beats any of the remaining combinations (Plu et al., 2015).

6. *Perform statistical significance testing* Considering the limited size and representativeness of the current evaluation datasets for the full complexity of disambiguation, we need solid methods and detailed analyses to understand the significance of our results. While at the moment empirical testing has an essential role in our field, statistical significance testing is often ignored or misused (Dror et al., 2018). Performing proper statistical significance testing could be very beneficial to assess whether a certain experimental result of a system is coincidental or not, whether its performance is optimized for certain cases more than others, and to which extent can that system be trusted to perform well on an unseen dataset in the future. In that sense, significance testing is a promising tool for better understanding and performance on the tail. Unfortunately, it was also not performed for the results obtained in this thesis, which is a problem in retrospective. Hopefully, future research will improve upon this situation and present significant improvements on the tail cases, in accordance with common practices in the field of statistics.

7.3 FUTURE RESEARCH DIRECTIONS

While I managed to address a set of questions regarding the tackling of long-tail identities in NLP, the research done so far is by no means complete. Instead, it might open more questions than it answers. The long-tail identities are a multifaceted challenge, and we are only starting to understand and measure their impact and complexity. There is a long list of future research directions provoked by this thesis, which constitute my agenda for subsequent investigations after this PhD cycle is concluded. These research directions can be organized along the following themes: 1. engineering of systems 2. novel tasks 3. a broader vision for the long tail. I now discuss each of these in more detail and with concrete scientific topics that I would pursue further.

7.3.1 *Engineering of systems*

1. *How to build systems that perform well on the tail?* It is still not entirely clear how to build systems that perform well on the tail. One option would be to ensure that the tail cases receive (much) higher weight than head cases in loss functions of probabilistic systems. A complementary idea would be to enhance these probabilistic systems with systematic usage and reasoning over extended knowledge. While this thesis provides the first steps in this direction by describing the head-tail distinction and by providing avenues for obtaining rich and diverse knowledge, future research should

proceed with investigations and engineering of these ideas into robust, high-performing systems on the tail.

2. *How to optimally combine the different types of background knowledge?* I discussed that certain knowledge types are consistently not integrated in existing entity linking systems. I also provided ways to obtain representative examples of these knowledge types, and showed their impact on the task of establishing identity of long-tail entities. It remains beyond the scope of this thesis to investigate optimal ways of combining these knowledge types. This challenge is especially relevant for the knowledge pieces which cannot be fed natively into the state-of-the-art probabilistic algorithms, such as information about the publishing time or the author of a document, or other documents that are essential or helpful to interpret the current document. Combining symbolic and probabilistic methods or information is a hot challenge of the modern day AI.

3. *What defines a 'good/useful' representation of an entity?* In our profiling work, I explained that our choice of properties was guided either by: a) properties generally studied in social psychology, or b) properties that can be found both in text and in knowledge bases. In a recent research, Amazon Alexa addresses this challenge from a different perspective and use case, by estimating the most relevant properties per entity type based on their demand in query logs (Hopkinson et al., 2018). The question of which properties (and values) can constitute a meaningful profile is rather complex, and it deserves more attention in subsequent works.

7.3.2 *Novel tasks*

4. *What is needed in order to shift the focus of our community towards long-tail and contextual cases?* The awareness in the semantic NLP community that certain kinds of cases are systematically easy for systems, whereas others are much more challenging, or even beyond their current capabilities, has been slowly increasing. The most straightforward strategy to make the long-tail cases more central in the current NLP research would probably entail better integration of these in current evaluation, either by: 1. increasing the prominence of tail cases in existing datasets, e.g., by deliberately creating long-tail datasets; 2. increasing their relevance in evaluation scores, e.g., by using macro F1-scores. Evaluating on long-tail cases would then increase the incentive for systems to address these issues, since that would directly correspond to their accuracy, and hence, publishability.

5. *How to enable and evaluate web-of-data-wide entity linking?* Whereas LOTUS allows access to much more knowledge about many more entities, there are various challenges that follow once the access has been improved. Considering that the entity linking task so far has been (almost) exclusively tested on DBpedia and other Wikipedia derivatives, it is non-trivial to adjust most EL systems to link to different knowledge bases, and especially

to a set of knowledge bases that use a different vocabulary, cover different domains, etc. Perhaps even more challenging is the evaluation: how does one evaluate a number of knowledge bases simultaneously in a reliable/-consistent way? This reopens many philosophical questions about identity which are out of the scope of this thesis, such as, which representation of a real-world entity is optimal? Is the Wikidata representation of John Travolta better than its IMDB or Musicbrainz representation? Intuitively, and keeping in mind that the goal of entity linking is to provide access to relevant and informative external background knowledge about an entity, choosing the best representation should be guided by the purpose of the linking and the context of the textual data.

A different (and less ambitious) approach towards web-of-data entity linking would be find a "sufficiently good" candidate for a given surface form, rather than a globally optimal one. This is common for tasks where the recall is hard to measure. In this sense, the task can be open-ended and post-validated, e.g., by crowd workers. Alternatively, it can be posed as a learning to rank task with a large, but finite, set of candidates for each entity, where the ranking within this set by the system is compared against the gold data via customary retrieval evaluation metrics, like Mean Average Precision and Mean reciprocal rank (Schütze et al., 2008).

6. *How can one capture non-factual beliefs?* Proceeding with the discussion in chapter 6, it remains a question to which extent can the human beliefs/expectations be substantiated by facts. Expectations that do not correspond to the facts (as captured by a knowledge base) pose an insurmountable limitation for our profiling machines. Filling this gap would require re-thinking and complementing the knowledge used by the profiler with knowledge or examples that capture non-factual social and cognitive expectations. Potentially these beliefs could be obtained through crowdsourcing tasks, but expertise and experiences from cognitive sciences should also be consulted to help answer this question.

7.3.3 *A broader vision for the long tail*

7. *What is the tail in other NLP tasks and how relevant is it?* I have provided a description of the head and the tail in the entity linking task, and observed that it does predict system performance consistently. How can this definition be transferred to other tasks? It seems that this transfer would be easiest for the most similar tasks, especially other reference tasks such as entity and event coreference. Once that is done, a follow-up investigation should explore to which extent is this head-tail distinction relevant for those other tasks.

8. *Explainable NLP?* Semantic NLP, and NLP in general, tends to employ statistical algorithms that behave as black boxes and operate on symbols that are not interpreted. For instance, in EL it is common to decide whether an

entity is a NIL or not based on a single scalar parameter called confidence threshold that is compared against the confidence score of each entity candidate as produced by the algorithm. Such examples are very prominent in NLP, but there are two grave interlinked problems regarding this kind of decisions within our NLP systems: such decisions are not transparent, and consequently, it is difficult to improve their functionality in an informed way. In this thesis, I propose algorithms based on interpreted knowledge rather than symbols that are not interpreted. Namely, by making use of the structured knowledge in the Semantic web, and building explicit profiles on top of it, as well as the knowledge classification theory, I made a step further towards explainable NLP techniques that can explain their decisions. For instance, in the profiling experiment in section 6, the system is able to explain why it concluded that entity A and B are the same, whereas C and D are not, based on their incomplete set of property values, enriched with background knowledge. I believe that there is vast room for improvement towards designing and implementing transparent and explainable NLP systems.

9. *How to build natively curious machines?* Long-tail instances suffer from knowledge scarcity ("hunger for knowledge"), which must be addressed in order to establish their identity correctly. One way to obtain this missing knowledge would be to mimic the human manner of curious, intrinsically motivated seeking for new needed knowledge. Our curiosity-driven search aims to advance our knowledge frontier and thus makes us selective, e.g., it tells us which document to read, or which page to attend to with more detail. This curiosity could be led by encountering unexpected/anomalous facts, by remaining ambiguity, or the needs of a downstream task.

   In a broader sense, this curiosity would be an essential skill within future computational knowledge bases that, alike the human mind, can simultaneously perform information extraction, anomaly detection, profiling, etc., and gradually push forward their knowledge boundary by being curious/hungry for obtaining new knowledge.

BIBLIOGRAPHY

Abourbih, Jonathan Alexander, Alan Bundy, and Fiona McNeill (2010). "Using Linked Data for Semi-Automatic Guesstimation." In: *AAAI Spring Symposium: Linked Data Meets AI*.

Abreu, Carla and Eugénio Oliveira (2018). "FEUP at SemEval-2018 Task 5: An Experimental Study of a Question Answering System." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics.

Adel, Heike, Benjamin Roth, and Hinrich Schütze (2016). "Comparing convolutional neural networks to traditional models for slot filling." In: *arXiv preprint arXiv:1603.05157*.

Adida, B., I. Herman, M. Sporny, and M. Birbeck (2012). *RDFa 1.1 Primer*. en. Tech. rep. http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/: World Wide Web Consortium.

Agirre, Eneko, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers (2010). "SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain." In: *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*. Uppsala, Sweden: Association for Computational Linguistics, pp. 75–80. URL: http://www.aclweb.org/anthology/S10-1013.

Akbari, Mohammad and Tat-Seng Chua (2017). "Leveraging Behavioral Factorization and Prior Knowledge for Community Discovery and Profiling." In: *Proceedings of the ACM Conference on Web Search and Data Mining*, pp. 71–79.

Angeli, Gabor, Julie Tibshirani, Jean Wu, and Christopher D Manning (2014). "Combining distant and partial supervision for relation extraction." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1556–1567.

Antoniou, Grigoris, Paul Groth, Frank van Harmelen, and Rinke Hoekstra (2012). *A Semantic Web Primer*. 3rd Edition. The MIT Press.

Ashmore, Richard D and Frances K Del Boca (1981). "Conceptual approaches to stereotypes and stereotyping." In: *Cognitive processes in stereotyping and intergroup behavior* 1, p. 35.

Aydilek, Ibrahim Berkan and Ahmet Arslan (2013). "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm." In: *Information Sciences* 233, pp. 25–35.

Bagga, Amit and Breck Baldwin (1998). "Algorithms for scoring coreference chains." In: *The first international conference on language resources and evaluation workshop on linguistics coreference*. Vol. 1. Granada, pp. 563–566.

Baker, Collin F, Charles J Fillmore, and Beau Cronin (2003). "The structure of the FrameNet database." In: *International Journal of Lexicography* 16.3, pp. 281–296.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (2013). "Abstract Meaning Representation for Sembanking." In: *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pp. 178–186.

Beek, Wouter and Laurens Rietveld (2015). "Frank: Algorithmic Access to the LOD Cloud." In: *Proceedings of the ESWC Developers Workshop 2015*.

Beek, Wouter, Laurens Rietveld, Hamid R Bazoobandi, Jan Wielemaker, and Stefan Schlobach (2014). "LOD laundromat: a uniform way of publishing other people's dirty data." In: *ISWC 2014*, pp. 213–228.

Beek, Wouter, Laurens Rietveld, Filip Ilievski, and Stefan Schlobach (2017). "LOD Lab: Scalable Linked Data Processing." In: *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering*. Lecture notes of summer school. Springer International Publishing, pp. 124–155.

Beek, Wouter, Filip Ilievski, Jeremy Debattista, Stefan Schlobach, and Jan Wielemaker (2018). "Literally better: Analyzing and improving the quality of literals." In: *Semantic Web Journal (SWJ)* 9.1, pp. 131–150. DOI: 10.3233/SW-170288. URL: https://doi.org/10.3233/SW-170288.

Bejan, Cosmin Adrian and Sanda Harabagiu (2010). "Unsupervised event coreference resolution with rich linguistic features." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1412–1422.

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio (2010). "Theano: A CPU and GPU math compiler in Python." In: *Proc. 9th Python in Science Conf*, pp. 1–7.

Bethard, Steven, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen (2015). "SemEval-2015 Task 6: Clinical TempEval." In: *SemEval@ NAACL-HLT*, pp. 806–814.

Boguraev, Branimir, James Pustejovsky, Rie Ando, and Marc Verhagen (2007). "TimeBank evolution as a community resource for TimeML parsing." In: *Language Resources and Evaluation* 41.1, pp. 91–115.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In: *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko (2013). "Translating embeddings for modeling multi-relational data." In: *Advances in neural information processing systems*, pp. 2787–2795.

Buil-Aranda, Carlos, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche (2013). "SPARQL Web-Querying Infrastructure: Ready for Action?" In: *ISWC 2013*.

Cano, Amparo E, Giuseppe Rizzo, Andrea Varga, Matthew Rowe, Stankovic Milan, and Aba-Sah Dadzie (2014). "Making Sense of Microposts (#Microp-

osts2014) Named Entity Extraction & Linking Challenge." In: *4th International Workshop on Making Sense of Microposts*. #Microposts.

Carpuat, Marine, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger (2013). "Sensespotting: Never let your parallel data tie you to an old domain." In: *Proceedings of the Association for Computational Linguistics (ACL)*. Citeseer.

Carreras, Xavier and Lluís Màrquez (2004). "Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004." In: chap. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. URL: http://aclweb.org/anthology/W04-2412.

Caselli, Tommaso and Roser Morante (2016). "VUACLTL at SemEval 2016 Task 12: A CRF Pipeline to Clinical TempEval." In: *Proceedings of SemEval*, pp. 1241–1247.

Che, Wanxiang and Ting Liu (2010). "Jointly modeling WSD and SRL with Markov logic." In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Association for Computational Linguistics, pp. 161–169.

Chen, Zheng and Heng Ji (2009). "Graph-based event coreference resolution." In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, pp. 54–57.

Cheng, Gong, Weiyi Ge, and Yuzhong Qu (2008). "Falcons: Searching and Browsing Entities on the Semantic Web." In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China: ACM, pp. 1101–1102. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367676. URL: http://doi.acm.org/10.1145/1367497.1367676.

Cheng, Xiao and Dan Roth (2013). "Relational inference for wikification." In: *Urbana* 51.61801, pp. 16–58.

Christophides, Vassilis, Vasilis Efthymiou, and Kostas Stefanidis (2015). *Entity Resolution in the Web of Data*. Morgan & Claypool Publishers.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1, pp. 37–46.

Corney, David, Dyaa Albakour, Miguel Martinez, and Samir Moussa (2016). "What do a Million News Articles Look like?" In: *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016*. Pp. 42–47. URL: http://ceur-ws.org/Vol-1568/paper8.pdf.

Corral, Álvaro, Gemma Boleda, and Ramon Ferrer-i Cancho (2015). "Zipf's law for word frequencies: Word forms versus lemmas in long texts." In: *PloS one* 10.7, e0129031.

Cybulska, Agata and Piek Vossen (2014). "Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 4545–4552.

Cybulska, Agata and Piek Vossen (2015). “" Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates.” In: *Int. J. Comput. Linguistics Appl.* 6.2, pp. 11–27.

Cyganiak, Richard, David Wood, and Markus Lanthaler (2014). *RDF 1.1 Concepts and Abstract Syntax*. W3C.

Daiber, Joachim, Max Jakob, Chris Hokamp, and Pablo N Mendes (2013). “Improving efficiency and accuracy in multilingual entity extraction.” In: *Proceedings of SEMANTiCS*. ACM, pp. 121–124.

Daume III, Hal (2007). “Frustratingly Easy Domain Adaptation.” In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 256–263. URL: http://www.aclweb.org/anthology/P07-1033.

Davis, Mark and Ken Whistler (2012). *Unicode Normalization Forms*. Unicode Consortium. URL: http://www.unicode.org/reports/tr15/tr15-37.html.

Derczynski, Leon, Kalina Bontcheva, and Ian Roberts (2016). “Broad twitter corpus: A diverse named entity recognition resource.” In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1169–1179.

Dijker, Anton J and Willem Koomen (1996). “Stereotyping and attitudinal effects under time pressure.” In: *European Journal of Social Psychology* 26.1, pp. 61–74.

Ding, Li, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs (2004). “Swoogle: A Search and Metadata Engine for the Semantic Web.” In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. CIKM ’04. Washington, D.C., USA: ACM, pp. 652–659. ISBN: 1-58113-874-1. DOI: 10.1145/1031171.1031289. URL: http://doi.acm.org/10.1145/1031171.1031289.

Dong, Xin, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang (2014). “Knowledge vault: A web-scale approach to probabilistic knowledge fusion.” In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 601–610.

Dror, Rotem, Gili Baumer, Segev Shlomov, and Roi Reichart (2018). “The hitchhiker’s guide to testing statistical significance in natural language processing.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1383–1392.

Elbassuoni, S., M. Ramanath, R. Schenkel, and G. Weikum (2010). “Searching RDF graphs with SPARQL and keywords.” In: 16?24.

Esquivel, José, Dyaa Albakour, Miguel Martinez, David Corney, and Samir Moussa (2017). “On the Long-Tail Entities in News.” In: *European Conference on Information Retrieval*, pp. 691–697.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd*. Vol. 96. 34, pp. 226–231.

Farid, Mina, Ihab F Ilyas, Steven Euijong Whang, and Cong Yu (2016). "LONLIES: estimating property values for long tail entities." In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, pp. 1125–1128.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Feyznia, Azam, Mohsen Kahani, and Fattane Zarrinkalam (2014). "COLINA: A Method for Ranking SPARQL Query Results Through Content and Link Analysis." In: *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*. ISWC-PD'14. Riva del Garda, Italy: CEUR-WS.org, pp. 273–276. URL: http://dl.acm.org/citation.cfm?id=2878453.2878522.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016). "Domain-adversarial training of neural networks." In: *Journal of Machine Learning Research* 17.59, pp. 1–35.

Gautam, Chandan and Vadlamani Ravi (2015). "Data imputation via evolutionary computation, clustering and a neural network." In: *Neurocomputing* 156, pp. 134–142.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press.

Graus, David, Tom Kenter, Marc Bron, Edgar Meij, Maarten De Rijke, et al. (2012). "Context-Based Entity Linking-University of Amsterdam at TAC 2012." In: *TAC*.

Graus, David, Manos Tsagkias, Wouter Weerkamp, Edgar Meij, and Maarten de Rijke (2016). "Dynamic collective entity representations for entity ranking." In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pp. 595–604.

Graus, David, Daan Odijk, and Maarten de Rijke (2018). "The birth of collective memories: Analyzing emerging entities in text streams." In: *Journal of the Association for Information Science and Technology* 69.6, pp. 773–786.

Grice, H. P. (1975). "Logic and Conversation." In: *Syntax and Semantics: Vol. 3: Speech Acts*. New York: Academic Press, pp. 41–58. URL: http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf.

Guha, Anupam, Mohit Iyyer, Danny Bouman, Jordan Boyd-Graber, and Jordan Boyd (2015). "Removing the Training Wheels: A Coreference Dataset that Entertains Humans and Challenges Computers." In: *North American Association for Computational Linguistics (NAACL)*.

Guu, Kelvin, John Miller, and Percy Liang (2015). "Traversing Knowledge Graphs in Vector Space." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 318–327. URL: http://aclweb.org/anthology/D15-1038.

Halpin, Harry (2012). *Social semantics: The search for meaning on the web*. Vol. 13. Springer Science & Business Media.

Heino, Erkki, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho, and Eero Hyvönen (2017). "Named entity linking in a complex domain: Case second world war history." In: *International Conference on Language, Data and Knowledge*. Springer, pp. 120–133.

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). "Teaching machines to read and comprehend." In: *Advances in Neural Information Processing Systems*, pp. 1693–1701.

Hewlett, Daniel, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot (2016). "WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1535–1545. DOI: 10.18653/v1/P16-1145. URL: http://aclweb.org/anthology/P16-1145.

Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum (2011). "Robust disambiguation of named entities in text." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 782–792.

Hoffart, Johannes, Yasemin Altun, and Gerhard Weikum (2014). "Discovering emerging entities with ambiguous names." In: *Proceedings of the 23rd international conference on World wide web*. ACM, pp. 385–396.

Hogan, Aidan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker (2011). "Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 9.4. {JWS} special issue on Semantic Search, pp. 365 –401. ISSN: 1570-8268. DOI: http://dx.doi.org/10.1016/j.websem.2011.06.004. URL: http://www.sciencedirect.com/science/article/pii/S1570826811000473.

Hogan, Aidan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker (2012). "An Empirical Survey of Linked Data Conformance." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 14, pp. 14–44.

Hong, Yu, Di Lu, Dian Yu, Xiaoman Pan, Xiaobin Wang, Yadong Chen, Lifu Huang, and Heng Ji (2015). "Rpi blender tac-kbp2015 system description." In: *Proc. Text Analysis Conference (TAC2015)*.

Hong, Yu, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer (2016). "Building a Cross-document Event-Event Relation Corpus." In: *LAW X*.

Hopkinson, Andrew, Amit Gurdasani, Dave Palfrey, and Arpit Mittal (2018). "Demand-Weighted Completeness Prediction for a Knowledge Base." In: *arXiv preprint arXiv:1804.11109*.

Hovy, Dirk and Anders Søgaard (2015). "Tagging performance correlates with author age." In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 483–488.

Hovy, Eduard, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot (2013). "Events are not simple: Identity, non-identity, and quasi-identity." In: *Workshop on Events: Definition, Detection, Coreference, and Representation*, pp. 21–28.

Hulpuş, Ioana, Narumol Prangnawarat, and Conor Hayes (2015). "Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation." In: *Proceedings of the International Semantic Web Conference (ISWC)*. Springer, pp. 442–457.

Hume, David (1738). *A Treatise of Human Nature*. Oxford University Press.

Ichinose, Shiori, Ichiro Kobayashi, Michiaki Iwazume, and Kouji Tanaka (2014). "Semantic Technology: Third Joint International Conference, JIST 2013, Seoul, South Korea, November 28–30, 2013, Revised Selected Papers." In: ed. by Wooju Kim, Ying Ding, and Hong-Gee Kim. Cham: Springer International Publishing. Chap. Ranking the Results of DBpedia Retrieval with SPARQL Query, pp. 306–319. ISBN: 978-3-319-06826-8. DOI: 10.1007/978-3-319-06826-8_23. URL: http://dx.doi.org/10.1007/978-3-319-06826-8_23.

Ilievski, Filip, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2015). "LOTUS: Linked Open Text UnleaShed." In: *Proceedings of the Consuming Linked Data (COLD) workshop*.

Ilievski, Filip, Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy (2016a). "Context-enhanced Adaptive Entity Linking." In: *LREC*.

Ilievski, Filip, Wouter Beek, Marieke van Erp, Laurens Rietveld, and Stefan Schlobach (2016b). "LOTUS: Adaptive Text Search for Big Linked Data." In: *European Semantic Web Conference (ESWC) 2016*. Springer International Publishing, pp. 470–485.

Ilievski, Filip, Marten Postma, and Piek Vossen (2016c). "Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text?" In: *The 26th International Conference on Computational Linguistics (COLING 2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1180–1191. URL: http://aclweb.org/anthology/C16-1112.

Ilievski, Filip, Piek Vossen, and Marieke van Erp (2017). "Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking." In: *International Conference on Language, Data and Knowledge*. Springer, Cham, pp. 143–149.

Ilievski, Filip, Piek Vossen, and Stefan Schlobach (2018). "Systematic Study of Long Tail Phenomena in Entity Linking." In: *The 27th International Conference on Computational Linguistics (COLING 2018)*.

Ilievski, Filip, Eduard Hovy, Qizhe Xie, and Piek Vossen (2018). "The Profiling Machine: Active Generalization over Knowledge." In: *ArXiv e-prints*. arXiv: 1810.00782 [cs.AI].

Iyyer, Mohit, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III (2014). "A Neural Network for Factoid Question Answering over Paragraphs." In: *EMNLP*, pp. 633–644.

Ji, Guoliang, Kang Liu, Shizhu He, and Jun Zhao (2016). "Knowledge Graph Completion with Adaptive Sparse Transfer Matrix." In: *AAAI*, pp. 985–991.

Ji, Heng and Ralph Grishman (2011). "Knowledge base population: Successful approaches and challenges." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 1148–1158.

Ji, Heng, Joel Nothman, Ben Hachey, and Radu Florian (2015). "Overview of TAC-KBP2015 tri-lingual entity discovery and linking." In: *Proceedings of the Eighth Text Analysis Conference (TAC2015)*.

Jiang, Jing and ChengXiang Zhai (2007). "Instance weighting for domain adaptation in NLP." In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Vol. 7, pp. 264–271.

Jurgens, David (2013). "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships." In: *ICWSM* 13.13, pp. 273–282.

Jurgens, David and Mohammad Taher Pilehvar (2016). "Semeval-2016 task 14: Semantic taxonomy enrichment." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1092–1102.

Jussim, Lee, Jarret T Crawford, and Rachel S Rubinstein (2015). "Stereotype (in) accuracy in perceptions of groups and individuals." In: *Current Directions in Psychological Science* 24.6, pp. 490–497.

Kahneman, Daniel and Amos Tversky (1979). "Prospect theory: An analysis of decision under risk." In: *Econometrica: Journal of the econometric society*, pp. 263–291.

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson, and Simon Kirby (2017). "Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication." In: *Cognition* 165, pp. 45–52.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization." In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kingsbury, Paul and Martha Palmer (2002). "From TreeBank to PropBank." In: *LREC*. Citeseer, pp. 1989–1993.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh (2009). "What's in Wikipedia? : mapping topics and conflict using socially annotated category structure." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, Pages 1509–1512.

Knuth, Magnus, Jens Lehmann, Dimitris Kontokostas, Thomas Steiner, and Harald Sack (2015). "The DBpedia Events Dataset." In: *International Semantic Web Conference (Posters & Demos)*.

Krippendorff, Klaus (1980). "Content analysis. Beverly Hills." In: *California: Sage Publications* 7, pp. l–84.

Kunda, Ziva (1999). *Social cognition: Making sense of people*. MIT press.

Lakshminarayan, Kamakshi, Steven A Harp, and Tariq Samad (1999). "Imputation of missing data in industrial databases." In: *Applied Intelligence* 11.3, pp. 259–275.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data." In: *Biometrics*, pp. 159–174.

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky (2012). "Joint entity and event coreference resolution across documents." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*. Association for Computational Linguistics, pp. 489–500.

Lei, Yuangui, Victoria Uren, and Enrico Motta (2006). "Semsearch: A search engine for the semantic web." In: *Managing Knowledge in a World of Networks*. Springer, pp. 238–245.

Lenzi, Valentina Bartalesi, Giovanni Moretti, and Rachele Sprugnoli (2012). "CAT: the CELCT Annotation Tool." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) 2012*. Istanbul, Turkey: European Language Resources Association (ELRA). URL: http://www.aclweb.org/anthology/L12-1072.

Ling, Xiao, Sameer Singh, and Daniel S Weld (2015a). "Design challenges for entity linking." In: *TACL* 3, pp. 315–328.

— (2015b). "Design Challenges for Entity Linking." In: *Transactions of the Association for Computational Linguistics* 3, pp. 315–28.

Liu, Yingchi and Quanzhi Li (2018). "NAI-SEA at SemEval-2018 Task 5: An Event Search System." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics.

Liu, Zhengzhong, Jun Araki, Eduard H Hovy, and Teruko Mitamura (2014). "Supervised Within-Document Event Coreference using Information Propagation." In: *LREC*, pp. 4539–4544.

Locke, John (1689). *An Essay Concerning Human Understanding*. Oxford University Press.

Lu, J. and V. Ng (2016). "Event Coreference Resolution with Multi-Pass Sieves." In: *LREC 2016*, pp. 3996–4003.

Luo, Xiaoqiang (2005). "On coreference resolution performance metrics." In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pp. 25–32.

MacLachlan, Gale and Ian Reid (1994). "Framing and interpretation." In:

Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews (2012). "Where Is This Tweet From? Inferring Home Locations of Twitter Users." In: *ICWSM* 12, pp. 511–514.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll (2004). "Finding predominant word senses in untagged text." In: *Proceedings of the 42nd An-*

nual Meeting on Association for Computational Linguistics (ACL 2004). Association for Computational Linguistics, p. 279.

Meij, Edgar, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke (2011). "Mapping queries to the Linking Open Data cloud: A case study using DBpedia." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 9.4, pp. 418–433.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son (2016). "MEANTIME, the NewsReader multilingual event and time corpus." In: *Proceedings of LREC2016*.

Mirza, Paramita, Fariz Darari, and Rahmad Mahendra (2018). "KOI at SemEval-2018 Task 5: Building Knowledge Graph of Incidents." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics.

Mitamura, Teruko, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel (2015a). "Event nugget annotation: Processes and issues." In: *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 66–76.

Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy (2015b). "Overview of TAC-KBP 2015 Event Nugget Track." In: *Text Analysis Conference*.

Mohammad, Saif and Graeme Hirst (2012). "Distributional Measures as Proxies for Semantic Relatedness." In: *CoRR* abs/1203.1.

Monahan, Sean, John Lehmann, Timothy Nyberg, Jesse Plymale, and Arnold Jung (2011). "Cross-Lingual Cross-Document Coreference with Entity Linking." In: *TAC*.

Moro, Andrea and Roberto Navigli (2015). "SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking." In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 288–297.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli (2014). "Entity linking meets word sense disambiguation: a unified approach." In: *TACL* 2, pp. 231–244.

Moussallem, Diego, Ricardo Usbeck, Michael Röeder, and Axel-Cyrille Ngonga Ngomo (2017). "MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach." In: *Proceedings of the Knowledge Capture Conference*. ACM, p. 9.

Mulay, Kunal and P. Sreenivasa Kumar (2011). "SPRING: Ranking the Results of SPARQL Queries on Linked Data." In: *Proceedings of the 17th International Conference on Management of Data*. COMAD '11. Bangalore, India: Computer Society of India, 12:1–12:10. URL: http://dl.acm.org/citation.cfm?id=2591338.2591350.

Nagy, István and Richárd Farkas (2012). "Person Attribute Extraction from the Textual Parts of Web Pages." In: *Acta Cybern.* 20.3, pp. 419–440.

Nakov, Preslav, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree (2015). "SemEval-2015 Task 3: Answer Selection in Community Question Answering." In: *Proceedings of the 9th International Work-*

shop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, pp. 269–281. DOI: 10.18653/v1/S15-2047. URL: http://aclweb.org/anthology/S15-2047.

Nakov, Preslav, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree (2016). "SemEval-2016 Task 3: Community Question Answering." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 525–545. DOI: 10.18653/v1/S16-1083. URL: http://aclweb.org/anthology/S16-1083.

Navarro, Gonzalo (2001). "A guided tour to approximate string matching." In: *ACM computing surveys (CSUR)* 33.1, pp. 31–88.

Navigli, Roberto and Simone Paolo Ponzetto (2012). "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." In: *Artificial Intelligence* 193, pp. 217–250.

Navigli, Roberto, David Jurgens, and Daniele Vannella (2013). "SemEval-2013 Task 12: Multilingual Word Sense Disambiguation." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 222–231. URL: http://www.aclweb.org/anthology/S13-2040.

Newman, Mark EJ (2005). "Power laws, Pareto distributions and Zipf's law." In: *Contemporary physics* 46.5, pp. 323–351.

Nguyen, Thien Huu, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi (2016a). "Joint Learning of Local and Global Features for Entity Linking via Neural Networks." In: *proceedings of COLING*.

Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng (2016b). "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset." In: *CoRR* abs/1611.09268. URL: http://arxiv.org/abs/1611.09268.

Nuzzolese, Andrea Giovanni, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli (2015). "Open knowledge extraction challenge." In: *Semantic Web Evaluation Challenge*. Springer, pp. 3–15.

O'Gorman, Tim, Kristin Wright-Bettner, and Martha Palmer (2016). "Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation." In: *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pp. 47–56.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999). "The PageRank citation ranking: bringing order to the web." In:

Palmer, Martha, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang (2001). "English Tasks: All-Words and Verb Lexical Sample." In: *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*. Toulouse, France: Association for Computa-

tional Linguistics, pp. 21–24. URL: http://www.aclweb.org/anthology/S01-1005.

Paulheim, Heiko and Christian Bizer (2014). "Improving the Quality of Linked Data Using Statistical Distributions." In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 10.2, pp. 63–86.

Pavlick, Ellie, Heng Ji, Xiaoman Pan, and Chris Callison-Burch (2016). "The Gun Violence Database: A new task and data set for NLP." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1018–1024. URL: http://aclweb.org/anthology/D16-1106.

Pearl, Judea (2009). *Causality*. Cambridge University Press.

Pearson, Ronald K (2006). "The problem of disguised missing data." In: *ACM SIGKDD Explorations Newsletter* 8.1, pp. 83–92.

Peng, Haoruo, Yangqiu Song, and Dan Roth (2016). "Event Detection and Co-reference with Minimal Supervision." In: *EMNLP*, pp. 392–402.

Phillips, A. and M. Davis (2009). *Tags for Identifying Languages*. RFC. URL: http://www.rfc-editor.org/info/rfc5646.

Piccinno, Francesco and Paolo Ferragina (2014). "From TagME to WAT: a new entity annotator." In: *Proceedings of the first international workshop on Entity recognition & disambiguation*. ACM, pp. 55–62.

Plu, Julien, Giuseppe Rizzo, and Raphaël Troncy (2015). "Revealing entities from textual documents using a hybrid approach." In: *3rd International Workshop on NLP & DBpedia, Bethlehem, Pennsylvania, USA*. Citeseer.

Postma, Marten, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen (2016a). "Addressing the MFS Bias in WSD systems." In: *Proceedings of LREC 2016*. Portorož, Slovenia: ELRA. ISBN: 978-2-9517408-9-1.

Postma, Marten, Filip Ilievski, Piek Vossen, and Marieke van Erp (2016b). "Moving away from semantic overfitting in disambiguation datasets." In: *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. Austin, TX: Association for Computational Linguistics, pp. 17–21. URL: http://aclweb.org/anthology/W16-6004.

Postma, Marten, Filip Ilievski, and Piek Vossen (2018). "SemEval-2018 Task 5: Counting Events and Participants in the Long Tail." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer (2007). "SemEval-2007 Task-17: English Lexical Sample, SRL and All Words." In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 87–92. URL: http://www.aclweb.org/anthology/S/S07/S07-1016.pdf.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (2011). "CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes." In: *Proceedings of the Fifteenth*

Conference on Computational Natural Language Learning (CoNLL 20111)-Shared Task. Association for Computational Linguistics, pp. 1–27.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (2012). "CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes." In: *Joint Conference on Empirical Methods on Natural Language Processing (EMNLP) and on Natural Language Learning (EMNLP-CoNLL 2012)-Shared Task*. Association for Computational Linguistics, pp. 1–40.

Preiss, Judita (2006). "A detailed comparison of WSD systems: an analysis of the system answers for the Senseval-2 English all words task." In: *Natural Language Engineering* 12.03, pp. 209–228.

Pustejovsky, James and Marc Verhagen (2009). "SemEval-2010 task 13: evaluating events, time expressions, and temporal relations (TempEval-2)." In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics, pp. 112–116.

Radford, Will, Ben Hachey, Matthew Honnibal, Joel Nothman, and James R Curran (2011). "Naıve but effective NIL clustering baselines–CMCRC at TAC 2011." In: *Proceedings of Text Analysis Conference (TAC 2011)*.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100, 000+ Questions for Machine Comprehension of Text." In: *CoRR* abs/1606.05250. URL: http://arxiv.org/abs/1606.05250.

Recasens, Marta and Eduard Hovy (2011). "BLANC: Implementing the rand index for coreference evaluation." In: *Natural Language Engineering,17, (4)*, pp. 485–510.

Recasens, Marta, Eduard Hovy, and M Antònia Martí (2011). "Identity, non-identity, and near-identity: Addressing the complexity of coreference." In: *Lingua* 121.6, pp. 1138–1152.

Richardson, Matthew, Christopher JC Burges, and Erin Renshaw (2013). "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text." In: *EMNLP*. Vol. 3, p. 4.

Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M Marlin (2013). "Relation Extraction with Matrix Factorization and Universal Schemas." In: *HLT-NAACL*, pp. 74–84.

Rietveld, Laurens, Wouter Beek, and Stefan Schlobach (2015a). "LOD lab: Experiments at LOD scale." In: *International Semantic Web Conference*. Springer, pp. 339–355.

— (2015b). "LOD Lab: Experiments at LOD Scale." In: *The Semantic Web – ISWC 2015*. Springer International Publishing, pp. 339–355.

Ristoski, Petar, Christian Bizer, and Heiko Paulheim (2015). "Mining the Web of Linked Data with RapidMiner." In: *Journal of Web Semantics* 35.3, pp. 142–151.

Rizzo, Giuseppe and Raphaël Troncy (2012). "NERD: a framework for unifying named entity recognition and disambiguation extraction tools." In: *Proceedings of EACL 2012*, pp. 73–76.

Rizzo, Giuseppe, Cano Amparo E, Bianca Pereira, and Andrea Varga (2015). "Making sense of Microposts (#Microposts2015) named entity recognition & linking challenge." In: 5^{th} *International Workshop on Making Sense of Microposts*. #Microposts.

Röder, Michael, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both (2014). "N^3-A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format." In: *LREC*, pp. 3529–3533.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

Snyder, Benjamin and Martha Palmer (2004). "The English All-Words Task." In: *SensEval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Ed. by Rada Mihalcea and Phil Edmonds. Barcelona, Spain: Association for Computational Linguistics, pp. 41–43. URL: http://www.aclweb.org/anthology/W/W04/W04-0811.pdf.

Socher, Richard, Danqi Chen, Christopher D Manning, and Andrew Ng (2013). "Reasoning with neural tensor networks for knowledge base completion." In: *Advances in neural information processing systems*, pp. 926–934.

Srivastava, Nitish, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Steinmetz, Nadine, Magnus Knuth, and Harald Sack (2013). "Statistical Analyses of Named Entity Disambiguation Benchmarks." In: *Proceedings of NLP & DBpedia 2013 workshop in conjunction with 12th International Semantic Web Conference (ISWC2013), CEUR Workshop Proceedings*.

Stone, GC, NL Gage, and GS Leavitt (1957). "Two kinds of accuracy in predicting another's responses." In: *The Journal of Social Psychology* 45.2, pp. 245–254.

Thalhammer, Andreas and Achim Rettinger (2016). "PageRank on Wikipedia: towards general importance scores for entities." In: *International Semantic Web Conference*. Springer, pp. 227–240.

Thomson, Judith Jarvis (1997). "People and their bodies." In: *Reading parfit* 202, pp. 202–05.

Tjong Kim Sang, Erik F. and Fien De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In: *Proceedings of CoNLL-2003*. Edmonton, Canada, pp. 142–147.

Tran, Thanh, Haofen Wang, and Peter Haase (2009). "Hermes: Data Web search on a pay-as-you-go integration infrastructure." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3. The Web of Data, pp. 189 – 203. ISSN: 1570-8268. DOI: http://dx.doi.org/10.1016/j.websem.2009.07.001. URL: http://www.sciencedirect.com/science/article/pii/S1570826809000213.

Tristram, Felix, Sebastian Walter, Philipp Cimiano, and Christina Unger (2015). "Weasel: a machine learning based approach to entity linking combining

different features." In: *Proceedings of 3th International Workshop on NLP and DBpedia, ISWC 2015*.

Tummarello, Giovanni, Renaud Delbru, and Eyal Oren (2007). "Sindice.Com: Weaving the Open Linked Data." In: *Proceedings ISWC'07/ASWC'07*. Busan, Korea, pp. 552–565. ISBN: 3-540-76297-3, 978-3-540-76297-3.

Usbeck, Ricardo, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both (2014). "AGDISTIS-graph-based disambiguation of named entities using linked data." In: *ISWC*. Springer, pp. 457–471.

Usbeck, Ricardo, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann (2015). "GERBIL: General Entity Annotator Benchmarking Framework." In: *24th International Conference on World Wide Web*. WWW.

Van Erp, Marieke, Filip Ilievski, Marco Rospocher, and Piek Vossen (2015). "Missing Mr. Brown and Buying an Abraham Lincoln-Dark Entities and DBpedia." In: *NLP-DBPEDIA@ ISWC*, pp. 81–86.

Van Erp, Marieke, Pablo N Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis (2016). "Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Vol. 5, p. 2016.

Van Herwegen, Joachim, Laurens De Vocht, Ruben Verborgh, Erik Mannens, and Rik Van de Walle (2015). "Substring filtering for low-cost Linked Data interfaces." In: *International Semantic Web Conference*. Springer, pp. 128–143.

Verborgh, Ruben, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle (2014). "Querying datasets on the web with high availability." In: *The Semantic Web–ISWC 2014*. Springer International Publishing, pp. 180–196.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A model-theoretic coreference scoring scheme." In: *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics, pp. 45–52.

Vossen, Piek (2018). "NewsReader at SemEval-2018 Task 5: Counting events by reasoning over event-centric-knowledge-graphs." In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA: Association for Computational Linguistics.

Vossen, Piek and Agata Cybulska (2016). "Identity and granularity of events in text." In: *Cicling 2016. Konya, Turkey*. URL: http://arxiv.org/abs/1704.04259.

Vossen, Piek, Ruben Izquierdo, and Atilla Görög (2013). "DutchSemCor: in quest of the ideal sense-tagged corpus." In: *Proceedings of Recent Advances in Natu-*

ral Language Processing (RANLP). INCOMA Ltd. Shoumen, Bulgaria, pp. 710–718.

Vossen, Piek, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers (2016). "Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news." In: *Knowledge-Based Systems* 110, pp. 60–85.

Vossen, Piek, Filip Ilievski, Marten Postma, and Roxane Segers (2018a). "Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vossen, Piek, Marten Postma, and Filip Ilievski (2018b). "ReferenceNet: a semantic-pragmatic network for capturing reference relations." In: *Global Wordnet Conference 2018, Singapore*.

Vrandečić, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase." In: *Communications of the ACM* 57.10, pp. 78–85.

Waitelonis, Jörg, Claudia Exeler, and Harald Sack (2015). "Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval." In: *Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC2015), CEUR Workshop Proceedings*.

Wang, Haofen, Qiaoling Liu, Thomas Penin, Linyun Fu, Lei Zhang, Thanh Tran, Yong Yu, and Yue Pan (2009). "Semplore: A scalable {IR} approach to search the Web of Data." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3. The Web of Data, pp. 177 –188. ISSN: 1570-8268. DOI: http://dx.doi.org/10.1016/j.websem.2009.08.001. URL: http://www.sciencedirect.com/science/article/pii/S1570826809000262.

Wang, Mengqiu, Noah A Smith, and Teruko Mitamura (2007). "What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA." In: *EMNLP-CoNLL*. Vol. 7, pp. 22–32.

Weston, Jason, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov (2015). "Towards AI-complete question answering: A set of prerequisite toy tasks." In: *arXiv preprint arXiv:1502.05698*.

Williams, Bernard A. O. (1957). "Personal Identity and Individuation." In: *Proceedings of the Aristotelian Society* 67.n/a, pp. 229–52.

Xie, Qizhe, Xuezhe Ma, Zihang Dai, and Eduard Hovy (2017). "An Interpretable Knowledge Transfer Model for Knowledge Base Completion." In: *Association for Computational Linguistics (ACL)*.

Yang, Yi, Wen-tau Yih, and Christopher Meek (2015). "WikiQA: A Challenge Dataset for Open-Domain Question Answering." In: *Proceedings of EMNLP*. Citeseer, pp. 2013–2018.

Yao, Limin, Sebastian Riedel, and Andrew McCallum (2013). "Universal schema for entity type prediction." In: *Proceedings of the 2013 workshop on automated KB construction*. ACM, pp. 79–84.

Yosef, Mohamed Amir, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum (2011). "Aida: An online tool for accurate disambiguation of named entities in text and tables." In: *Proceedings of the VLDB Endowment* 4.12, pp. 1450–1453.

Zheng, Jin G, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji (2015). "Entity linking for biomedical literature." In: *BMC medical informatics and decision making* 15.1, S4.

Zhong, Bei, Jin Liu, Yuanda Du, Yunlu Liaozheng, and Jiachen Pu (2016). "Extracting attributes of named entity from unstructured text with deep belief network." In: *International Journal of Database Theory and Application* 9.5, pp. 187–196.

Zipf, George (1935). *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, Mass.: M.I.T. Press.

Zwicklbauer, Stefan, Christin Seifert, and Michael Granitzer (2016). "DoSeR-A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings." In: *ISWC*. Springer, pp. 182–198.

SUMMARY

The digital era and the omnipresence of computer systems feature a huge amount of data on identities (profiles) of people, organizations, and other entities, in a digital format. This data largely consists of textual documents, such as news articles, encyclopedias, personal websites, books, and social media, thus transforming identity from a philosophical to a societal issue and motivating the need for robust computational tools that can determine entity identity in text. Determining the identity of the entities described in these documents is non-trivial, given their amount, the contextual dependency of these descriptions, the ambiguity and variance of language, and the interplays described in widely-accepted pragmatic laws of language, such as the Gricean maxims (Grice, 1975).

Today, it is well-understood how to determine identity of entities with low ambiguity and high popularity/frequency in communication (*the head*), as witnessed by the high accuracy scores in the standard Natural Language Processing (NLP) task of Entity Linking. It is unclear, however, how to interpret *long-tail* entities: each different and potentially very ambiguous, with low frequency/popularity, and scarce knowledge.

Expecting that computational systems that establish identity in text struggle with long-tail cases, this thesis investigated how the performance of NLP techniques for establishing the identity of long-tail cases can be improved through the use of background knowledge. It focused on five aspects of this challenge: description/definition, analysis, improvement of evaluation, enabling access to more knowledge, and building knowledge-intensive systems. Concretely, the research questions and corresponding findings of this thesis were:

- *How can the tail entities be distinguished from head entities?* Our experiments showed a positive dependency of system performance on frequency and popularity of entity instances, and a negative one with ambiguity of surface forms. Essentially, this confirms the intuition that system performance is largely based on head cases, and declines strongly on the tail.
- *Are the current evaluation datasets and metrics representative for the long-tail cases?* The commonly used datasets to evaluate disambiguation and reference NLP tasks lack representativeness, as they suffer from low ambiguity, low variance, high dominance, and limited temporal spread.
- *How can we improve the evaluation on the long-tail cases?* On a deliberately created task to evaluate tail instances, we observed very low accuracy of the participating systems. This shows that dealing with high ambiguity and not being able to rely on frequency biases, poses a great challenge for current NLP systems.
- *How can the knowledge on long-tail entities be accessed and enriched beyond DBpedia?* By creating LOTUS, we provided the Linked Open Data community

with the largest centralized text index and access point to the LOD Laundromat data collection. This allows EL systems to use the knowledge found among the billions of statements of the LOD Laundromat collection, thus essentially increasing their recall on the tail instances.

- *What is the added value of background knowledge models when establishing the identity of NIL entities?* Neural background knowledge models ("profiling machines") were built and applied in order to complete the partial profiles extracted from text and establish their identity. The evaluation of these machines on the task of establishing long-tail identity in text showed promising results when applied on top of automatically extracted information from text. We observed no clear patterns between the effectiveness of our profilers and the data ambiguity.

This dissertation thus provided novel insights into an under-explored and difficult NLP challenge: determining identity of long-tail entities in text, demonstrating that better evaluation and more extensive use of knowledge are promising directions forward. The topics covered and the skills employed stemmed from various AI fields: semantic NLP, semantic web, and neural networks, with links to linguistics and philosophy. Besides summarizing a range of learned lessons that are potentially applicable to a number of other disambiguation and reference tasks, this thesis also provoked a long list of future research directions.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`