SwissLink: High-Precision, Context-Free Entity Linking Exploiting Unambiguous Labels

Roman Prokofyev
Michael Luggen
Djellel Eddine Difallah
Philippe Cudré-Mauroux
{firstname.lastname}@unifr.ch
eXascale Infolab, University of Fribourg—Switzerland

ABSTRACT

Webpages are an abundant source of textual information with manually annotated entity links, and are often used as a source of training data for a wide variety of machine learning NLP tasks. However, manual annotations such as those found on Wikipedia are sparse, noisy, and biased towards popular entities. Existing entity linking systems deal with those issues by relying on simple statistics extracted from the data. While such statistics can effectively deal with noisy annotations, they introduce bias towards head entities and are ineffective for long tail (e.g., unpopular) entities. In this work, we first analyze statistical properties linked to manual annotations by studying a large annotated corpus composed of all English Wikipedia webpages, in addition to all pages from the CommonCrawl containing English Wikipedia annotations. We then propose and evaluate a series of entity linking approaches, with the explicit goal of creating highly-accurate (precision > 95%) and broad annotated corpuses for machine learning tasks. Our results show that our best approach achieves maximal-precision at usable recall levels, and outperforms both state-of-the-art entity-linking systems and human annotators.

KEYWORDS

Entity Linking, Manual annotations, Machine learning

ACM Reference format:
Roman Prokofyev, Michael Luggen, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2017. SwissLink: High-Precision, Context-Free Entity Linking Exploiting Unambiguous Labels. In *Proceedings of Semantics2017, Amsterdam, Netherlands, September 11–14, 2017,* 8 pages.
https://doi.org/10.1145/3132218.3132234

1 INTRODUCTION

A large fraction of the data available both on the Web and inside enterprises is composed of unstructured data, particularly textual data such as Web pages, emails, social media content, etc. By its very nature, textual data can be easily understood by humans in individual pieces, but represents a major challenge for large-scale, automated processing and understanding by machines. The process of automatically understanding textual data usually consists in multiple steps, where most steps corresponds to tasks that aim to mine structured information from the textual contents. One particularly important task in that context is Entity Linking, which aims to enhance textual data with structured elements (and thus make it understandable by machines).

Entity Linking (also known as named entity disambiguation) is the task of correctly linking entities appearing in a text to their representation in a given knowledge base. Entities in that context can describe real-world objects, persons, or concepts, much like entries in an encyclopedia. Entity linking has become an integral part of modern information retrieval systems and semantic search engines [1].

To perform entity linking, a system first needs to generate entity candidates for potential mentions of entities in text. Many state-of-the-art entity linking systems use Wikipedia as a foundation of their link generation algorithms [9, 16, 17]. There, Wikipedia is used to construct a database of *entities* and their corresponding *textual representations* (surface forms), in addition to any information that appears on a page i.e., other entities, other surface forms etc. For example, an entity "John_F._Kennedy"[1] has the following surface forms: {"Kennedy", "JFK",.., "John Kennedy"}.

Before a correct link to an entity can be provided for a given surface form in a text, it is often necessary to *disambiguate* the surface form at hand, that is, to determine the correct link from a set of potential candidates. Techniques to do so include the use of semantic networks extracted from a knowledge base [11], hierarchical topic models [12], graph-based approaches on linked data [4], or hybrid human-machine techniques [5].

Effective entity linking with unconstrained text is still a challenging problem today. Despite many years of research, automated state-of-the-art entity linking approaches, such as DBpedia Spotlight, produce many false linkings, especially when it comes to disambiguating the labels. The low precision of existing system is particularly problematic when the linked entities are used for training a classifier using machine learning techniques or when the linked entities are directly exposed to the users of a semantic or search application.When the precision is low, the performance of the trained models is poor and the users get unsatisfactory results, leading to low acceptance of the system in general.

[1] https://en.wikipedia.org/wiki/John_F._Kennedy

Semantics2017, September 11–14, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5296-3/17/09…$15.00
https://doi.org/10.1145/3132218.3132234

The main idea of modern entity linking system is that they aim to strike a balance between precision and recall, e.g. by optimizing their F1 score. However, since even state-of-the-art systems reporting high F1 scores yield precision values that are below human annotators, the produced output is far too noisy for machine learning applications requiring nearly perfect precision of the input data. Hence, current Entity Linking approaches are unfortunately not usable in many machine learning scenarios.

In this work, we propose instead a new entity linking approach with the goal of *achieving very high precision* (+95%) *without sabotaging recall* such that it can be used in many machine learning applications. Achieving such high precision requires very strict criteria when considering a named entity as a linkable candidate. Our approach consists in computing a set of highly specific labels extracted from crowd-annotated data, such as Wikipedia pages. In order not to penalize the recall too severely, we introduce a relaxed measure of ambiguity that has its roots in the prior probability of an entity. In contrast to other methods, we do not consider the surrounding context of named entities (since it often introduces uncertainty), but solely rely on statistics that we extract from Wikipedia and the Web. We experimentally prove the viability of our approach in Section 5.

In summary, the main contributions of our work are:

- A novel statistical method to produce high precision entity annotations based on highly specific labels for Big Data and machine learning scenarios;
- A study of crowd-inferred entity annotations from anchors gathered on the Web, and a classification of the errors they could induce;
- An extensive evaluation of the effectiveness of the proposed approaches with different parameters and thresholds on our parameters setting dataset;
- A comparision of our method against a human annotations and DBpedia Spotlight tuned for precision.
- A validation using an established entity linking framework on 9 commonly used datasets.

The rest of the paper is organized as follows: We start with an overview of related work in the areas of entity linking, entity typing and language models in Section 2 below. Section 3 provides the motivation for our approach, describes the knowledge base we use, and introduces our new entity linking methods. Section 5 describes our training and test collections, our experimental setting, and presents the results of empirical evaluation comparing various combinations of our entity linking approaches. Finally, we conclude and discuss future work in Section 6.

2 RELATED WORK

In the following, we review the relevant works on Entity Linking as well as highlight the differences in our approach.

Entity linking can be divided into three subtasks: i) mention detection, ii) link generation, and iii) disambiguation [13]. *Mention detection* is an Natural language processing (NLP) topic mostly, as it identifies named entities (e.g., persons, organizations, locations) inside textual contents, using either rule-based systems analyzing sentence structures, co-occurrence statistics [19] or through more robust techniques such as string matching [2, 6].

Once a named entity has been detected, the link generation identifies its relevant counterpart in the knowledge base (whenever it exists).

Recent work integrates a series of probabilities in a combined generative model for entity linking [10]. Information taken into account for the linking process includes prior probabilities of entities (the likelihood of a surface form to refer to a certain entity), surface form statistics (probability of a given spelling being used) and some textual context in relation to their connectedness in the Knowledge Base (e.g., probability of being used with other words or entities in the same phrase). In our approach, we additionally take into account statistical information relating to the presence of surface forms in large-scale online sources and the distribution of their referred entities.

Before a correct link can be provided, it might be necessary to *disambiguate* the entity at hand, that is, to distinguish the correct link from a set of ambiguous candidates. Entity disambiguation techniques include methods that leverage semantic networks extracted from a knowledge base [11], hierarchical topic models [12], graph-based approaches on linked data [4], and hybrid human-machine techniques [5].

In Wikify! [16], the reported techniques yielding the best results exploit prior-probabilities of entities in Wikipedia. Obviously, this approach performs well on popular entities only. To mitigate this bias, prior-probabilities of entities can be balanced with graph-relatedness metrics among all entities for each document [17].

Further features extracted from the corpus such as the Redirects or Disambiguation pages in Wikipedia can be incorporated to improve the results [3]. Our approach includes all the above features as a baseline to build on.

Entity linking techniques are often context and corpus-dependent and have to be adapted to properly work in different contexts. The Tagme disambiguation system [7], for instance, is a recent system that is based on anchors and entities from Wikipedia, but focuses on short textual snippets by finding the collective agreement between the candidate entities. Meij *et al.* proposed in a similar context a machine learning solution for micro-blog posts which uses a plethora of features categorized into common *concept features*, *n-gram features* such as prior-probabilities, and finally some micro-blogging specific *tweet features* [14]. The input in that context is always a complete message with some closed context (e.g, the full text of a tweet.) Our method is also capable of finding entities in short documents as it is by nature indifferent to the length of a document and as it does not rely on any contextual information.

A number of systems are readily available online for entity linking, taking plain text as input and providing links (annotations) to the counterpart entities in the knowledge base as output (e.g., DBpedia Spotlight [15], Babelfy [18]). The performance of such systems is however far from being perfect; DBpedia Spotlight reports for example an F1 score of 56% ($P \approx 0.67$, $R \approx 0.48$) [15] when annotating plain text.

Similarly to Babelfy, REL-RW [9] adopts a graph-based approach leveraging random walks, and computes the semantic similarities of entities and documents. Subsequently, it selects the entity with the maximum similarity score for a given mention. Unlike Babelfy,

Table 1: (left) Top five entity/label links on the Web. (right) Entity linked with the label 'Wikipedia'

Entity	Label	Count
<ISO_639:en>	*en*	3'064'718
<Web_browser>	*web browser*	381'623
<Roche_limit>	*Roche limit*	365'970
<Yo-yo>	*yo-yo*	365'969
<Centripetal_force>	*centripetal force*	365'968

Entity	Count
<Main_Page>	163'032
<Belt_buckle>	16'560
<Angelina_Jolie>	16'333
<Jakarta_Tourism_and_Culture_Office>	9'473
<Project_Runway>	6'196

Table 2: Entities Linked by Context (in bold is the label used in the hyperlink).

Context and Label	Entity
divided into three subgroups: ***East****, West, and South*	<East_Slavic_languages>
and the ***first*** *African American to hold the office*	<List_of_African-American_firsts>
actions of two groups: ***gun control*** *and*	<Gun_politics_in_the_United_States>
television personality, author, and ***candidate*** *for the*	<Donald_Trump_presidential_campaign,_2016>
and was ***inaugurated as president*** *on January 20, 2009*	<First_inauguration_of_Barack_Obama>

however, it can only disambiguate entities given correct mentions in a document.

The latest published approach for entity linking, PBoH [8], proposes a probabilistic graphical model to represent the problem and approximate inference techniques to generate the results. The inference techniques used in that context rely on three features similar to the ones in [10], namely, prior probabilities of entities, pairwise co-occurrences of entities within documents, and entity-contextual-word co-occurrence statistics.

In contrast to the previous approaches, we focus on the first stage of entity linking, that is, candidate selection. This step is often overlooked, since most of the approaches rely on a list of candidate-links ranked by prior-probability subject to errors, and bias (popularity of entities). In the following, we study and classify such errors and we introduce new methods that are context-free and focus on reducing the bias and the error in the annotations produced by crowd-annotators and found on the Web.

3 MOTIVATION

High-precision entity linking is essential for a number of applications where even a small fraction of erroneous links are highly detrimental. One example is entity linking for interactive applications with humans. At Armasuisse[2], for instance, several data integration systems work by aggregating textual sources based on sets of predefined entities. The resulting integrated text is then reviewed by human analysts, whose work is to summarize and analyze the results. In that context, even a very small fraction of errors in the links can lead to situations where the analysts cannot work properly and waste hours trying to synthesize various pieces of information that should not have been linked in the first place, and that creates an increasing overhead for the analysts.

Another very timely application of high-precision linking is machine learning; increasingly today, annotated (i.e., linked) text is used for training all kinds of predictive models (see section 2). In this context, achieving a very high precision (+95%) while delivering a broad enough training corpus calls for very strict criteria when considering a named entity as a linkable candidate. In that sense, we have a pretty different goal than state-of-the-art entity linking approaches, which aim to strike an ideal balance between precision and recall, e.g., by optimizing their F1 score. While state-of-the-art approaches try to disambiguate entities by leveraging their textual context, including both local (surrounding text) and global (other entities/words in the text) contexts, we focus our attention on the labels that are unambiguous by themselves when referring a certain entity, for example: *"JFK airport"*. Hence, automatically inferring this information is at the center of this work. This means, however, that some ambiguous labels will be ultimately excluded from the dataset and never linked. We try to mitigate this problem with a method that relaxes strict ambiguity property of a label in Section 4 using a set of heuristics.

3.1 What Links to Wikipedia?

Existing Entity Linking systems rely on large collections of human-annotated textual data, where words and phrases in documents are annotated with links to entities from some knowledge base. However, these annotations are noisy, partly because these documents were annotate in order to be read by other humans, rather than by machines.

In order to understand how annotations are used on the Web, we crawled all entity links found on two large datasets, by processing the CommonCrawl [3] and the Wikipedia dumps [4]. The output of our processing is a list of all words and phrases that were used as anchors in hyperlinks pointing to English Wikipedia pages, together with the URIs of these pages. From our result analysis, we can distinguish the following cases when linking to Wikipedia:

- *Link by the most common label.* From Table 1, we observe that the top entity-label pairs are correctly linking labels to their corresponding Wikipedia pages. In general, these

[2] http://www.ar.admin.ch/

[3] http://commoncrawl.org/ (November 2015).

[4] https://dumps.wikimedia.org/ (November 2015).

Table 3: (left) frequent labels that were linked to a few different entities by reference. (right) Normal frequent words that were linked to many different entities with a reference. $C(l_i)$ denotes the number of occurrences of a label in a plain text of a corpus, $C(l_i, *)$ denotes the number of occurrences of a label as an anchor for some entity. And Uri Count is the number of entities with such label.

Label	Uri Count	$C(l_i)$	$C(l_i, *)$
customers	2	47K	184
school year	2	24K	131
courage	1	12K	106
lunch	2	10K	98

Label	Uri Count	$C(l_i)$	$C(l_i, *)$
song	164	257K	2.3K
group	140	228K	2.2K
book	151	128K	1K
system	133	252K	0.7K

annotations are the most useful for entity linking, and their frequency is a good indication of such an association, albeit the label itself can be ambiguous (see "Yo-yo" disambiguation page for other possible uses [5]).

- *Link by context.* In some cases, annotators create a link to an entity using a somewhat inappropriate label, although the context of the sentence clearly refers to that entity. See Table 2 for examples of such annotations.
- *Link by reference.* On the Web, many links to Wikipedia emanate from short labels as the annotator is pointing to Wikipedia for more information on a particular subject, but using a label such as *"About"* or *"Wikipedia"*. If not properly removed, the frequency of such labels can yield wrong links. Other examples of such links include labels like "book", "song", "game", "character", etc., linking to specific books, songs, etc (Table 3).
- *Erroneous link.* By examining the links and their labels, we notice many wrong associations of entity-label pairs that cannot be rooted to the previous cases. Examples of such cases include incorrect label boundaries, such as anchor "prime minister of" pointing to <Minister_President_of_Prussia>, or "Oregon" pointing to <University_of_Oregon>; both examples were incorrect even taking into account their respective contexts, and thus are not instances of a Link by Context.

3.2 Prior Probability and Ambiguous labels

As overviewed in Section 2, most Entity Linking systems consider a Prior Probability score expressed as the conditional probability of a link $P(link|label)$ derived from the raw counts of the occurrences of the links, the labels and label-link pairs found in the training corpus.

To derive a high-precision entity linking method that does not take any contextual information into account, we need to base our decision solely on this collected data. Prior probability is in this context insufficient, as it does not necessarily capture the ambiguity of a label; nor does it guarantee that the entity with the highest score is the correct one (as reported in the literature and demonstrated in section 3.1). In fact, it does not take into account the different categories of links that we identified previously, and thus can be subject to errors induced by humans or by the relatively broad ambiguity of a label. For instance, the label *"Wikipedia"* has a relatively high probability of referring to the entity <Belt_buckle>. Conversely, *"London"* would always link to the capital of the United Kingdom and not to the writer <Jack_London>.

In the following section, we introduce statistical methods for identifying and linking highly specific labels. These methods will be based on the observations that we made in this section and on the understanding of how crowd-annotators create Wikipedia annotations.

[5] https://en.wikipedia.org/wiki/Yo-yo_(disambiguation)

4 ENTITY LINKING METHODS

Our approach to Entity Linking consists in identifying sets of highly specific labels collected from a large dataset of crowd-annotated data. This step is completed in a batch mode, which allows us to perform *just-in-time* linking by scanning through a document, looking for such labels and thus detecting and linking the matching named entities. In this section, we explore a set of approaches and ambiguity metrics that will help us construct dictionaries of such specific labels.

Notations. Let E be the set of all entities, and L the set of English labels (strings) that have been used at least once as an anchor text for some entity. Let $C(e_i)$ be the number of anchors pointing to entity $e_i \in E$, $C(l_i)$ be the number of occurrences of label $l_i \in L$ in the corpus, and $C(l_i, e_j)$ the number of anchors pointing to entity e_j using l_i as anchor text. Note that: $\forall l_i \in L, \forall e_j \in E, C(l_i) \geq C(l_i, e_j)$.

Problem Formulation. The problem we tackle is the following: given an arbitrary textual document I_D as input, identify all named entities substrings $\{l_1, .., l_k\}$ and link them to their respective entities. Effectively, our methods will return as output a set of entity-label pairs $O_D = \{(l_1, e_z), ..., (l_k, e_x)\}$.

4.1 1-Entity-Labels

First, we explore the extreme case of labels referring to a single entity in the entire corpus, i.e., labels $l_i \in L$ satisfying:

$$\exists! e_i, C(l_i, e_j) > 0 \quad (1)$$

This method allows us to remove all ambiguous labels that were directly observed in the corpus. However, besides its extremely low Recall, this method is insufficient to obtain specific entity labels, for the following reasons:

- 1-Entity-labels with a low count do not present statistically significant information to decide on their ambiguity (as ambiguous cases can be rare for some entities and hence unobservable in small corpuses or for unpopular entities);

Table 4: (left) Labels with many disambiguations and low ratios. (right) Specific labels with low ratios

Label	Uri Count	$C(l_i)$	$C(l_i, *)$
Canada	530	91K	75K
Paris	263	48K	39K
railway station	981	40K	8K
Clinton	119	7K	1K

Label	Uri Count	$C(l_i)$	$C(l_i, *)$
sea snail	2	18K	17K
Rotten Tomatoes	1	9K	9K
iTunes	2	11K	7K
DC Comics	3	9K	5K

- Links by context (see section 3.1) will likely yield erroneous 1-Entity labels, e.g., in an article about "Amilcar Compound"[6] and in the following sentence: "government would block the company", the phrase "would block" links to "Paul-Marie Pons", an engineer who created the Pons Plan which restructured the French auto-industry after the Second World War. . Such links often use common phrases that would not be linked to the entity outside of the context.

4.2 Label-Entity Ratio

To address the apparent issues of the previous method, we introduce some post-filtering on the labels. As such, for every 1-Entity-label l_i we filter out a label if the following condition is met:

$$\frac{C(l_i)}{C(l_i, e_j)} > threshold \tag{2}$$

The intuition behind our filtering technique is that the labels with unusually high counts as found by string matching are likely to be ambiguous, especially for those with a significantly higher count than the actual entity they link to. For instance, this method can allow us to filter out labels that were linked by context.

4.3 Percentile-Ratio method

The previous methods take extreme measures when it comes to isolating specific labels, thus severely penalizing Recall. In the following, we introduce a relaxed measure of ambiguity that draws its roots in the Prior Probability of the candidates. In essence, we are aiming to identify those labels that are specific for certain entities, but which link to more than one entity due to erroneous linking.

For a given label l_i, the distribution $C(l_i, e_j)$ of all entities that the label links to gives an indication of the label ambiguity, e.g., a skewed distribution indicates a low ambiguity label followed by a tail of potentially erroneous annotations. Note that this indication can also be biased by the popularity of those links. For instance, "Moscow, Russia" appears orders of magnitude more often in Wikipedia than "Moscow, Indiana" or "Moscow, Idaho". Other examples of labels with skewed distributions include "Canada" and "Paris" (Table 4), which would be filtered out completely based on just the number of entities they link to.

We derive a measure that strikes a balance between these two observations by introducing a percentile cutoff on the entity count, thus keeping all labels that map only to a single entity after the percentile cutoff. Formally, for each label l_i, mapping to a set of entities E_{l_i}, we idenfity a reduced set of links as follows:

$$U_{l_i} = \{e_i \in E_{l_i} : C(e_i) \geq \alpha \times \sum_{k}^{|E_{l_i}|} C(e_k)\} \tag{3}$$

Where $\alpha \in (0, 1]$ is the percentile threshold imposed for the cutoff. Finally, we consider only the labels with $|U_{l_i}| = 1$ as being non-ambiguous, and we generate a link for them. Note that the number of specific labels we identify through the Percentile-Ratio method is necessarily greater or equal than the number of labels obtained from the ratio-based methods. It is exactly equal if $\alpha = 1$.

We empirically test and compare the above techniques on different input corpora in Section 5.

5 EXPERIMENTAL EVALUATION

In this section, we experimentally evaluate the effectiveness of our methods. We start by focusing on a test collection of Wikipedia articles that we manually annotated in order to better understand the performance of our various components and their parameters. Subsequently, we compare against a large range of entity linking methods and across several datasets.

5.1 Experimental Setup

Preprocessing. Our methods are based on statistics collected from Wikipedia dumps. In the following, we consider the Wikipedia dump from 2015-06-02. To preprocess the dump, we use the *Wikipedia Extractor*[7] library that converts raw Wikipedia pages into HTML documents. We customized the library to produce plain-text documents with links instead of HTML documents. In addition, we excluded all pages that belonged to special Wikipedia categories, such as "Disambiguation" or "Requests for adminship" using a list of stop URIs[8].

Wikipedia test collection. In order to evaluate, parametrize and compare the results of our methods, we decided to create our own collection of Wikipedia pages thoroughly annotated with entity links. We use this collection primarily to understand how various parameters should be set and how they influence the final quality of the results. Moreover, it allows us to test the widely adopted assumption about the correctness of the links created by the crowd on Wikipedia. In this process, We randomly selected 30 pages from Wikipedia and annotated each of them with Wikipedia entities found inside the pages. In detail, we first automatically identified candidate entities using all existing labels in Wikipedia, and then we examined and annotated each of those cases manually. The total number of valid entities found in these articles is 2908. The

[6] https://en.wikipedia.org/wiki/Amilcar_Compound

[7] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[8] http://uimr.deri.ie/sites/StopUris/

P@MI | R@MI P@MA | R@MA	ACE2004		AIDA/CoNLL-Complete		AQUAINT		DBpedia Spotlight		Microposts2014-Test		MSNBC		N3-Reuters-128		N3-RSS-500		OKE 2015 evaluation	
AIDA	11.23	6.86	49.42	37.20	9.82	3.71	67.39	9.39	63.90	14.09	30.41	18.07	54.26	25.34	63.43	31.40	79.88	39.46
	42.01	40.78	45.85	35.31	10.60	3.68	32.16	9.45	51.97	47.62	32.18	17.92	40.55	23.80	43.40	31.40	76.42	39.95
Babelfy	63.96	41.18	64.64	39.99	78.39	33.43	53.33	9.70	71.74	23.25	73.44	50.33	64.56	28.98	64.83	30.60	81.46	40.36
	75.06	62.29	59.33	36.33	76.86	32.94	33.07	9.77	58.04	52.65	68.61	46.27	47.50	26.42	43.90	30.60	77.05	40.91
DBpedia Spotlight	77.95	49.67	68.51	40.17	83.67	45.80	90.80	23.94	83.78	34.95	72.55	34.67	75.70	27.61	71.97	30.30	85.38	33.43
	83.71	68.48	69.38	40.41	83.67	45.75	65.01	24.13	69.19	61.17	74.18	37.36	52.47	25.84	43.70	30.30	76.57	33.36
Dexter	86.72	36.27	68.54	28.90	87.25	36.73	87.50	16.97	84.13	26.59	77.18	24.90	74.82	23.30	77.42	24.0	91.49	32.38
	84.68	59.42	67.96	28.94	87.84	37.35	56.55	18.22	63.96	56.75	75.83	27.83	51.25	23.45	39.30	24.0	83.61	33.88
Entity classifier	55.74	42.81	51.69	38.09	81.63	27.51	85.45	14.24	66.20	29.94	50.79	38.96	44.61	31.48	35.81	30.80	58.33	22.14
	68.21	63.49	50.08	37.13	76.67	26.67	43.58	13.55	63.68	57.56	49.34	38.48	40.62	30.53	34.40	30.80	52.13	22.57
NERD-ML	71.43	45.75	62.8	36.87	77.73	46.08	79.31	41.82	66.43	37.5	68.41	43.78	60.28	28.98	63.64	25.90	82.48	46.08
	79.62	66.57	59.61	35.62	78.24	45.52	69.76	45.48	67.15	61.18	68.78	45.45	46.15	26.55	37.20	25.90	78.23	46.55
TagMe 2	86.08	54.58	65.60	39.54	79.31	63.27	77.60	58.79	72.57	49.28	76.71	44.98	74.30	30.23	71.10	34.20	82.45	44.58
	88.26	70.98	65.82	37.06	78.66	61.41	74.38	60.61	76.38	70.46	75.59	45.34	54.79	28.25	46.80	34.20	82.11	46.94
WAT	77.16	49.67	68.66	50.70	**88.70**	36.73	80.0	9.70	78.90	20.54	81.01	45.11	78.91	36.14	68.91	31.70	81.02	40.51
	82.91	66.26	67.98	50.05	88.0	36.05	35.78	10.65	57.83	51.70	79.88	41.60	61.41	35.99	43.70	31.70	75.09	41.31
SwissLink	**90.91**	13.07	**83.66**	7.42	85.0	14.03	**94.29**	10.0	**100.0**	5.97	**98.43**	16.73	**90.62**	6.59	**90.72**	17.60	**97.60**	18.37
	66.58	42.95	47.17	6.78	70.67	13.98	42.53	11.26	47.30	44.36	99.27	17.47	28.91	6.26	32.70	17.60	68.98	20.86

Table 5: Micro and macro Precision and Recall results reported by Gerbil on 9 datasets on Swisslink. To put our SwissLink into perspective we also report the others systems results which were, unlike our system not tuned for precision. In bold, the best Precision at micro is highlighted.

selected articles, along with the annotations, are available online for reproducibility purposes[9].

Evaluation Metrics. We evaluate the entity linking methods at hand using standard metrics relevant to our use-case (high-precision linking) i.e., Precision and Recall. Both measures will be reported as i) Micro-Average (@MI): aggregated across mentions, and ii) Macro-Average (@MA): aggregated across documents.

Evaluation Framework. For our evaluations, we adopt the Gerbil testbench [21], a well-established framework for entity linking. In the context of Gerbil, the task we are solving is referred to as A2KB, that is: identifying and linking entities in text (while D2KB is the task of disambiguating entities given the correct entity mentions). Since our techniques are context-free, both A2KB and D2KB setups are relevant.

However, we note that our methods tend to extract more entities from the datasets than what is available in the ground-truth, and most of the missing annotation turn out to be correct when inspecting the ground truth. This is a well-known problem of non-iterative evaluation campaign and should be dealt-with in a continuous evaluation campaign [20], which unfortunately is not possible at this point for Gerbil. Since this has a direct impact on the precision (as Gerbil treats those cases as false positives), we report the D2KB task results instead as they more fairly reflect the quality of our method. Note that, however, we do not use the correct entity mentions provided by the D2KB task, and instead extract the mentions automatically.

5.2 Parameters Setting Experiment

In order to better grasp the performance of our various approaches and to adjust the parameters of the methods we introduce in section 4, we start by varying the ratio and the percentile thresholds from 0 to 100 and 90% to 99.9% respectively. In addtion, we compare against a number of baselines:

- DBpedia Spotlight: were in turn we vary the internal confidence parameter that is sensitive to "topical pertinence and contextual ambiguity" [15], in the 100%-70% confidence range;
- DBpedia Labels: the canonical representations of these entities are used as unambiguous labels, and;
- Manual Annotations: these are the annotations created by the human editors in Wikipedia.

Figure 1 shows the effect of the ratio and percentile parameters on our methods. Specifically, increasing the ratio from 1 up to 5 increases the recall without affecting significantly the precision (>95%) across the different percentiles used. The highest percentiles (e.g., 99.9%), on the other hand, seem to be more conservative and start loosing in recall abruptly when increasing the ratio, as opposed to lower percentiles.

The manual annotations have a recall of 37% (1076 total entities), because the articles are never fully annotated[10], and a high precision of 94%, diminished because of incorrect annotations.

[9] https://github.com/XI-lab/Wikipedia30

[10] Annotations in Wikipedia articles are not intended to be complete, as this reduces their readability by humans.

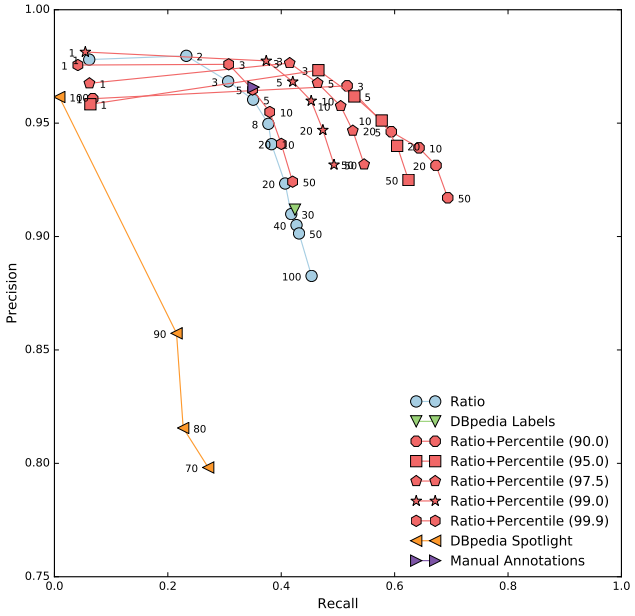

Figure 1: The Precision and Recall of entity linking methods on the Wikipedia test collection. Increasing the ratio threshold up to "50" (label in front of the teal triangles/circles) improves the recall in general while keeping the precision in the 90% band. DBpedia Spotlight confidence level (label in front of pink circles) increases the Precision at the expense of a high drop in Recall.

With DBpedia Spotlight, a 100% confidence yields a precision of 96%, with a recall of 1%. Reducing the confidence level causes a steep decline in precision in exchange for an increasing recall. Conversely, DBpedia Labels result in a recall of 56% for a maximum precision of 92%, which is lower than all our methods. Note that this method is a variant of our 1-entity Link, since each entity has exactly 1 label (see Section 4.1).

Discussion and Takeaways. By inspecting the above results in detail, we observe that when increasing the ratio in particular we are allowing more ambiguous labels to be introduced; this has a direct impact on precision (black circles in Figure 1. The percentile is nicely balancing this effect by separating the ambiguity from the popularity of the entities.

In general, we observe that the Percentile-Ratio method with 99-Percentile and 10-Ratio, strikes in our context a good balance between high-precision results (>95%) and reasonable recall (45%, 1309 entities). Thus we pick these parameters in the reminder of our evaluations and we refer to this setup as *SwissLink*.

5.3 Gerbil Experiments

In order to validate our results on commonly used datasets in the Entity Linking community, we decided to use Gerbil. We deployed our method configured with the Percentile-Ratio Ratio_Threshold=10 and 99-Percentile (as it was the best performing setup in our experiments) through a webservice compatible with Gerbil's API.

We then proceeded to an evaluation on all available datasets.[11]. For the purpose of putting our results into context (there is no possibility to tune the other systems for precision), we also compared to other available systems [12].

We report in Table 5 results pertaining to entity linking methods available on Gerbil to put our results into context. Please note that unlike SwissLink, they are tuned for F1 and can not be tuned for high precision. We can differentiate two types of metrics depending on the utilized aggregation method:

- Micro-average (MI), which computes Precision and Recall across all documents;
- Macro-average (MA), that first computes Precision and Recall for each document, and then takes an average.

Since MA results report document-level measures, SwissLink performs less competitively, due to the large discrepancy in document ambiguity levels. Specifically, if a document contains mainly ambiguous labels, our MA precision and recall are pushed to the lower scores.

Thus, we highlight the MI Precision and Recall, as they measure the total quality of the links produced.

As we observe, we achieve high scores with a P@MI ranging between 83.66% and 100%. The recall is in general lower than other methods as expected, though we reach reasonable levels in our context according to the use-cases we introduce in Section 3.

6 CONCLUSIONS

In this work, we have explored the task of high-precision entity linking and proposed a number of different approaches to link entities in such a way.

Our linking methods are based on the identification of highly specific labels, i.e. textual forms of entities that have a distinct entity attribution irrespective of their textual context. We proposed a number of methods to find such unambiguous labels that leverage statistics on user-created links in a large textual corpus, e.g., Wikipedia. We experimentally proved that the linking precision of our best-performing methods is comparable or even higher than the precision of user-created links for our task. Hence, we believe that our method can be applied to any not-annotated textual corpus to produce high-precision annotations. Moreover, since our approach is context-free, it can be used as an initial step in other entity linking systems to generate better candidates in place of the more traditional Prior-Probability. Finally, we made our method publicly available[13] for reproducibility purposes.

7 ACKNOWLEDGEMENT

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 683253/GraphInt).

REFERENCES

[1] R. Blanco, G. Ottaviano, and E. Meij. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 179–188. ACM, 2015.

[2] M. Ciaramita and Y. Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 594–602, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[3] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 708–716, 2007.

[4] P. Cudré-Mauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. idMesh: Graph-based disambiguation of linked data. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 591–600, New York, NY, USA, 2009. ACM.

[5] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM.

[6] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Communications ACM*, 51(12):68–74, Dec. 2008.

[7] P. Ferragina and U. Scaiella. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA, 2010. ACM.

[8] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann. Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 927–938, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[9] Z. Guo and D. Barbosa. Robust Entity Linking via Random Walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 499–508, New York, NY, USA, 2014. ACM.

[10] X. Han and L. Sun. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 945–954, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[11] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 215–224, New York, NY, USA, 2009. ACM.

[12] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1037–1045, New York, NY, USA, 2011. ACM.

[13] E. Meij, K. Balog, and D. Odijk. Entity linking and retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1127–1127, New York, NY, USA, 2013. ACM.

[14] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 563–572, New York, NY, USA, 2012. ACM.

[15] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[16] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.

[17] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[18] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2, 2014.

[19] R. Prokofyev, G. Demartini, and P. Cudré-Mauroux. Effective named entity recognition for idiosyncratic web collections. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 397–408, New York, NY, USA, 2014. ACM.

[20] A. Tonon, G. Demartini, and P. Cudré-Mauroux. Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval*, 18(5):445–472, Oct. 2015.

[21] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, et al. Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1133–1143. International World Wide Web Conferences Steering Committee, 2015.

[11] http://w3id.org/gerbil/experiment?id=201604300040

[12] http://w3id.org/gerbil/experiment?id=201605010002

[13] https://github.com/eXascaleInfolab/kilogram