VoxEL: A Benchmark Dataset for Multilingual Entity Linking

Henry Rosales-Méndez, Aidan Hogan and Barbara Poblete

Millenium Institute for Foundational Research on Data
Department of Computer Science, University of Chile
{hrosales,ahogan,bpoblete}@dcc.uchile.cl

Abstract. The Entity Linking (EL) task identifies entity mentions in a text corpus and associates them with corresponding entities in a given knowledge base. While traditional EL approaches have largely focused on English texts, current trends are towards language-agnostic or otherwise multilingual approaches that can perform EL over texts in many languages. One of the obstacles to ongoing research on multilingual EL is a scarcity of annotated datasets with the same text in different languages. In this work we thus propose VoxEL: a manually-annotated gold standard for multilingual EL featuring the same text expressed in five European languages. We first motivate and describe the VoxEL dataset, using it to compare the behavior of state of the art EL (multilingual) systems. In particular, we compared language-specific EL systems across five different languages and contrasted these results against those obtained using machine translation to English. Overall, our results show an important gap in the performance of language-specific EL systems and English. In addition, machine translation appears to be a competitive alternative for multilingual EL that is worth exploring.

Keywords: Benchmark, Dataset, Multilingual Entity Linking
Resource type: Dataset
Permanent URL: https://dx.doi.org/10.6084/m9.figshare.6104759

1 Introduction

The Entity Linking (EL) task identifies entity mentions in a text corpus and associates them with corresponding entities in a Knowledge Base (KB). In this way, we can leverage the information of publicly available KBs about real-world entities to achieve a better understanding of their semantics and also of natural language. For instance, in the text *"in the world of pop music, there is Michael Jackson and there is everybody else"* quoted from The New York Times, we can link the mention *Michael Jackson* with its corresponding entry in, for example, the Wikidata KB [32] (`wd:Q2831`), or the DBpedia KB [15] (`dbr:Michael_Jackson`)[1] allowing us to leverage, thereafter, the information in the KB about this entity to support semantic search, relationship extraction,

[1] Throughout, we use prefixes according to `http://prefix.cc`.

text enrichment, entity summarization, or semantic annotation, amongst other applications.

One of the major driving forces for research on EL has been the development of a variety of ever-expanding KBs that describe a broad selection of notable entities covering various domains (e.g., Wikipedia, DBpedia, Freebase, YAGO, Wikidata). Hence, while traditional Named Entity Recognition (NER) tools focused on identifying mentions of entities of specific types in a text, EL further requires disambiguation of which entity in the KB is being spoken about; this remains a challenging problem. On the one hand, name variations – such as *"Michael Joseph Jackson"*, *"Jackson"*, *"The King of Pop"* – mean that the same KB entity may be referred to in a variety of ways by a given text. On the other hand, ambiguity – where the name *"Michael Jackson"* may refer to various (other) KB entities, such as a journalist (`wd:Q167877`), a football player (`wd:Q6831558`), an actor (`wd:Q6831554`), and more – means that an entity mention in a text may have several KB candidates associated with it.

Many research works have addressed these challenges of the EL task down through the years. Most of the early EL systems proposed in the literature were monolingual approaches focusing on texts written in one single language, in most cases English [16,12]. These approaches often use resources of a specific language, such as Part-Of-Speech taggers and WordNet[2], which prevent generalization or adaptation to other languages. Furthermore, most of the labelled datasets available for training and evaluating EL approaches were English only (e.g., AIDA/CoNLL [12], DBpedia Spotlight Corpus[16], KORE 50 [13]).

However, as the EL area has matured, more and more works have begun to focus on languages other than English, including multilingual approaches that are either language agnostic [6,5,8,20] – relying only on the language of labels available in the reference KB – or can be configured for multiple languages [19,27]. Recognizing this trend, a number of multilingual datasets for EL were released, such as for the 2013 TAC KBP challenge[3] and the 2015 SemEval Task 13 challenge[4]. Although such resources are valuable for multilingual EL research – where in previous work [26] we presented an evaluation of EL systems comparing two languages from the SemEval dataset – they have their limitations, key amongst which are their limited availability (participants only[5]), a narrow selection of languages, and differences in text and annotations across languages that makes it difficult to compare the performance in each language. More generally, the EL datasets available in multiple languages – and languages other than English – greatly lags behind what is available for English.

Contributions: In this paper, we propose the VoxEL dataset: a manually-annotated gold standard for EL considering five European languages: German,

[2] `https://wordnet.princeton.edu`; April 1st, 2018
[3] `https://tac.nist.gov/2013/KBP/`; April 1st, 2018
[4] `http://alt.qcri.org/semeval2015/task13/`; April 1st, 2018
[5] We have managed to acquire the SemEval dataset, but unfortunately we were not able to acquire the TAC–KBP dataset: our correspondence was not responded to.

English, Spanish, French and Italian selected. This dataset is based on an online source of multilingual news, where we select and annotated 15 corresponding news articles for these five languages (giving 45 articles in total). Additionally, we created two versions of VoxEL: a *strict* version where entities correspond to a restricted definition of entity, as a mention of a person, place or organization (based on traditional MUC/NER definitions), and a *relaxed* version where we considered a broader selection of mentions referring to entities described by Wikipedia. Based on the VoxEL dataset, using the GERBIL evaluation framework [31], we present results for various EL systems, allowing us to draw comparisons not only across systems, but also across languages. As an additional contribution, we compare the performance of EL systems configurable for a given language with the analogous results produced by applying state-of-the-art machine translation (Google translate) to English and then applying EL configured for English. Our findings show that there is a significant difference in the performance of native English EL systems and multilingual EL. English EL on average outperforms multilingual EL by 61.9%. Furthermore, machine translation in addition to English EL systems achieves very similar performance to multilingual EL.

2 Preliminaries

We first introduce some preliminaries relating to EL. Let E be a set of entity identifiers in a KB; these are typically IRIs, such as `wd:Q2831`, `dbr:Michael_Jackson`. Given an input text, the EL process can be conceptualized in terms of two main phases. First, Entity Recognition (ER) establishes a set of entity mentions M, where each of such mentions is typically considered as a string that refers to an entity, annotated with its start position in the input text, e.g., (37, *"Michael Jackon"*). Second, for each mention $m \in M$ recognized by the first phase, Entity Disambiguation (ED) attepts to establish a link between m and the corresponding identifier $e \in E$ for the KB entity to which it refers. The second disambiguation phase can be further broken down into a number of (typical) sub-tasks, described next:

Candidate entity generation: For each mention $m \in M$, this stage selects a subset of the most probable KB entities $E_m \subseteq E$ to which it may refer. There are two high-level approaches by which candidate entities are often generated. The first is a dictionary-based approach, which involves applying keyword or string matching between the mention m and the label of entities from E. The second is an NER-based approach, where traditional NER tools are used to identify entity mentions (potentially) independently of the KB.

Candidate entity ranking: This stage is where the final disambiguation is made: the candidate entities E_m for each mention m are ranked according to some measure indicating their likelihood of being the reference for m. The measures used for ranking each entity $e \in E_m$ may take into account features of the candidate entity e (e.g., centrality), features of the candidate link (m, e)

(e.g., string similarity), features involving e and candidates for neighbouring mentions E'_m (e.g., graph distance in the KB), and so forth. Ranking may take the form of an explicit metric that potentially combines several measures, or may be implicit in the use of machine-learning methods that classify candidates, or that compute an optimal assignment of links.

Unlinkable mention prediction: The target KBs considered by EL are often, by their nature, incomplete. In some applications, it may thus be useful to extract entity mentions from the input text that do not (yet) have a corresponding entity in the KB. These are sometimes referred to as *emerging entities*, are typically produced by NER candidate generation (rather than a dictionary approach), and are assigned a label such as *NIL* (Not In Lexicon).

It is important to note that while the above processes provide a functional overview of the operation of most EL systems, not all EL systems follow this linear sequence of steps. Most systems perform recognition first, and once the mentions are identified the disambiguation phase is initiated [16,19]. However, other approaches may instead apply a unified process, building models that create feedback between the recognition and disambiguation steps [7]. In any case, the output of the EL process will be a set of links of the form (m, e), where the mention m in the text is linked to the entity e in the KB, optionally annotated with a confidence score – often called a *support* – for the link.

3 Related Work

We now cover related works in the context of multilingual EL, first discussing approaches and systems, thereafter discussing available datasets.

3.1 Multilingual EL Systems

In theory, any EL system can be applied to any language; as we demonstrated in our previous work [26], even a system supporting only English components may still be able to correctly recognize and link the name of a person such as *Michael Jackson* in the text of another language, assuming the alphabet remains the same. Hence, the notion of a multilingual EL system can become blurred. For example language-agnostic systems – systems that require no linguistic components or resources specific to a language – can become multilingual simply by virtue of having a reference KB with labels in a different – or multiple different – language(s).

Here we thus focus on EL systems that have published evaluation results over texts from multiple languages, thus demonstrating proven multilingual capabilities. We summarize such systems in Table 1, where we provide details on the year of the main publication, the languages evaluated, as well as denoting whether or not a demo, source code or API is currently available.[6] As expected, a high-level inspection of the table shows that English is the most popularly-evaluated

[6] We presented an earlier version of such a table in previous work [26].

Table 1. Overview of multilingual EL approaches; the italicized approaches will be incorporated as part of our experiments

Name	Year	Evaluated Languages	Demo	Src	API
KIM [22]	2004	English, French, Spanish	✓	✗	✓
TagME [8]	2010	English, German, Dutch	✓	✗	✓
SDA [3]	2011	English, French	✗	✗	✗
ualberta [10]	2012	English, Chinese	✗	✗	✗
HITS [7]	2012	English, Spanish, Chinese	✗	✗	✗
THD [6]	2012	English, German, Dutch	✓	✓	✓
DBpedia Spotlight [16,5]	2013	English, Italian, Russian, Dutch, French, German, Spanish, Hungarian, Danish	✓	✓	✓
Wang-Tang [34]	2013	English, Chinese	✗	✗	✗
AGDISTIS [30,20]	2014	English, German, Spanish French, Italian, Japanese, Dutch	✓	✓	✓
Babelfy [19]	2014	English, Spanish, French German, Italian	✓	✗	✓
FREME [27]	2016	English, German	✗	✓	✓
WikiME [29]	2016	English, Spanish, French, Italian, Chinese, German, Thai, Arabic, Turkish, Tamil, Tagalog, Urdu, Hebrew	✓	✗	✗
FEL [21]	2017	English, Spanish, Chinese	✗	✓	✗
FOX [28]	2017	English, German, Spanish, French, Dutch	✓	✓	✓

(and thus we surmise supported) language, followed by European languages such as German, Spanish, French, Dutch and Italian. We also highlight the year of publication, where – with the exception of KIM [22] – most of the included multilingual EL approaches have emerged since 2011.

We will later conduct experiments using the GERBIL evaluation framework [31], which allows for invoking and integrating the results of a variety of public APIs for EL, generating results according to standard metrics in a consistent manner. Hence, in our later experiments, we shall only consider those systems with a working REST-API made available by the authors of the system. In addition, we will manually label our VoxEL system according to Wikipedia, with which other important KBs such as DBpedia, YAGO, Freebase, Wikidata, etc., can be linked; hence we only include systems that support such a KB linked with Wikipedia. With these criteria in mind, we select the following systems:

THD (2012) is based on three measures [6]: *most frequent senses*, which ranks candidates for a mention based on the Wikipedia Search API results for that mention; *co-occurrence*, which is a co-citation measure looking at how often candidate entities for different mentions are linked from the same paragraphs in Wikipedia; and *explicit semantic analysis*, which uses keyword similarity measures to relate mentions with a concept. These methods are language agnostic and applicable to different language versions of Wikipedia.

DBpedia Spotlight (2013) was first proposed to deal with English annotations [16], based on keyword and string matching functions ranked by a probabilistic model based on a variant of a TF–IDF measure. DBpedia Spotlight is largely language-agnostic, where an extended version later proposed by Daiber et al. [5] leverages the multilingual information of the Wikipedia and DBpedia KBs to support multiple languages.

TagME (2013) uses analyses of anchor texts in Wikipedia pages to perform EL [8]. The ranking stage is based primarily on two measures: *commonness*, which describes how often an anchor text is associated with a particular Wikipedia entity; and *relatedness*, which is a co-citation measure indicating how frequently candidate entities for different mentions are linked from the same Wikipedia article. TagME is language agnostic: it can take advantage of the Wikipedia Search API to apply the same conceptual process over different language versions of Wikipedia to support multilingual EL.

AGDISTIS (2014) assumes that recognition has been performed using another tool and thus, given a set of mentions, focuses on the disambiguation process [30]. The generation of candidates is mainly performed over indexes computed off-line from the target KB, where the ranking stage computes an assignment of links from related entities in a graph built over neighbouring candidate entities and mentions. A recent study [20] extends this approach, adding some new features to improve its behavior in multilingual scenarios.

Babelfy (2014) performs EL with respect to a custom multilingual KB BabelNet [7] constructed from Wikipedia and WordNet, using machine translation to bridge the gaps in information available for different language versions of Wikipedia [19]. Recognition is based on POS tagging for different languages, selecting candidate entities by string matching. Ranking is reduced to finding the densest subgraph that relates neighbouring entities and mentions.

FREME (2016) delegates the recognition of entities to the Stanford-NER tool, which is trained over the anchor texts of Wikipedia corpora in different languages. Candidate entities are generated by keyword search over local indexes, which are then ranked based on the number of matching anchor texts in Wikipedia linking to the corresponding article of the candidate entity [27].

With respect to FOX, note that while it meets all of our criteria, at the time or writing, we did not succeed in getting the API to run over VoxEL without error; hence we do not include this system.

[7] `http://babelnet.org/`; April 1st, 2018

Table 2. Survey of dataset for EL task. For multilingual datasets, the quantities shown refer to the English data available. We present metadata about the relaxed and strict version of our dataset by VoxEL$_R$ and VoxEL$_S$ respectively. (Abbreviations: $|D|$ number of documents, $|S|$ number of sentences, $|E|$ number of entities, **Mn** denotes that all entities were manually annotated.)

Dataset	$\|D\|$	$\|S\|$	$\|E\|$	**Mn**	**Languages**
AIDA/CoNLL-Complete [12]	1393	22,137	34,929	✓	EN
KORE50 [13]	50	50	144	✓	EN
IITB [14]	103	1,781	18,308	✓	EN
ACE2004 [23]	57	-	306	✗	EN
AQUAINT [23]	50	533	727	✗	EN
MSNBC [4]	20	668	747	✗	EN
DBpedia Spotlight [16]	10	58	331	✓	EN
N3-RSS 500 [24]	1	500	1000	✓	EN
Reuters 128 [24]	128	-	881	✓	EN
Wes2015 [33]	331	-	28,586	✓	EN
News-100 [24]	100	-	1656	✓	DE
Thibaudet [1]	1	3,807	2,980	✗	FR
Bergson [1]	1	4,280	380	✗	FR
SemEval 2015 Task 13 [18]	4	137	769	✓	EN,ES,IT
DBpedia Abstracts [2]	39,132	-	505,033	✗	DE,EN,ES,FR, IT,JA,NL
MEANTIME [17]	120	597	2,790	✗	EN,ES,IT,NL
VoxEL$_R$	15	94	674	✓	DE,EN,ES,FR,IT
VoxEL$_S$	15	94	204	✓	DE,EN,ES,FR,IT

3.2 Multilingual EL Datasets

In order to train and evaluate EL approaches, labelled datasets – annotated with the correct entity mentions and their respective KB links – are essential. In some cases these datasets are labelled manually, while in other cases labels can be derived from existing information, such as anchor texts. In Table 2 we survey the labelled datasets most frequently used by EL approaches (note that sentence counts were not available for some datasets).

We can see that the majority of datasets provide text in one language only – predominantly English – with the exceptions being as follows:

SemEval 2015 Task 13: is built over a biomedical, math, computer and social domain and is designed to support EL and WSD at the same time, containing annotations to Wikipedia, BabelNet and WordNet [18].

DBpedia Abstracts: provides a large-scale training and evaluation corpora based on the anchor texts extracted from the abstracts (first paragraph) of Wikipedia pages in seven languages [2].[8]

MEANTIME: consists of 120 news articles from WikiNews[9] with manual annotations of entities, events, temporal information and semantic roles [17].[10]

With respect to DBpedia Abstracts, while offering a very large multilingual corpus, the texts across different languages varies, as do the documents available; while such a dataset could be used to compare different systems for the same languages, it could not be used to compare the same systems for different languages. Furthermore, there are no guarantees possible for the completeness of the annotations since they are anchor texts/links extracted from Wikipedia; hence the dataset is best suited as a (very) large collection of positive examples – in a similar manner to how TagME [8] and FREME [27] use anchor texts – rather than an evaluation dataset since it cannot reliably identify false positives.

Unlike DBpedia Abstracts, the SemEval and MEANTIME datasets contain analogous documents translated to different languages. However, MEANTIME and SemEval 2015 Task 13 uses machine translation from English to obtain the text in the other languages. Despite the tremendous progress made in machine translation in recent years, translation errors are still a possibility, particularly where entity labels are involved. For example, while Google Translate would currently translate the Spanish sentence "*Vi la película ≪Jungla de cristal≫ ayer.*" to "*I saw the movie 'Crystal Jungle' yesterday.*", a human expert would rather translate it as "*I saw the movie 'Die Hard' yesterday.*". In previous work we conducted a comparison of EL systems for English and Spanish texts using the SemEval dataset; we refer the reader to [26] for more details, including results.

4 The VoxEL Dataset

In this section, we describe the VoxEL Dataset that we propose as a gold standard for EL involving five languages: German, English, Spanish, French and Italian. VoxEL is based on 15 news articles sourced from the VoxEurop[11] web-site: a European newsletter with the same news articles professionally translated to different languages. This source of text thus obviates the need for translation of texts to different languages, and facilitates the consistent identification and annotation of mentions (and their Wikipedia links) across languages. With VoxEL, we thus provide a high-quality resource with which to evaluate the behavior of EL systems across a variety of European languages.

[8] `http://wiki-link.nlp2rdf.org/abstracts/`; April 1st, 2018
[9] `https://en.wikinews.org/`; April 1st, 2018
[10] `http://www.newsreader-project.eu/results/data/wikinews/`; April 1st, 2018
[11] `http://www.voxeurop.eu/`; April 1st, 2018

While the VoxEurop newsletter is a valuable source of professionally translated text in several European languages, there are sometimes natural variations across languages that – although they preserve meaning – may change how the entities are mentioned. A common example is the use of pronouns rather than repeating a person's name to make the text more readable in a given language. Such variations would then lead to different entity annotations across languages, hindering comparability. Hence, in order to achieve the same number of sentences and annotations for each new (document), we applied small manual edits to homogenize the text (e.g., replacing a pronoun by a person's name). On the other hand, sentences that introduce new entities in one particular language, or that deviate too significantly across all languages, are eliminated.

With respect to the guidelines for labelled entity mentions, we take into consideration the lack of consensus about what is an "*entity*": some works conservatively consider only mentions of entities referring to fixed types such as person, organization and location as entities (similar to the traditional NER/TAC consensus on an entity) [25], while other authors note that a much more diverse set of entities are available in Wikipedia and related KBs for linking, and thus consider any noun-phrase mentioning an entity in Wikipedia to be a valid target for linking [25]. Furthermore, there is a lack of consensus on how overlapping entities – like *New York City Fire Department* – should be treated [25]; should *New York City* be annotated as a separate entity or should we only cover maximal entities? Rather than take a stance on such questions – which appear application dependant – we instead create two versions of the data: a *strict* version that considers only maximal entity mentions referring to persons, organizations and locations; and a *relaxed* version that considers any noun phrase mentioning a Wikipedia entity as a mention, including overlapping mentions where applicable. For example, by the sentence *"Michael Jackson is the dancer with the fanciest feet on the street"* we would only include in the strict version to "Michael Jackson", while in the relaxed version we would also include "dancer", "feet" and "street".

To create the annotation of mentions with corresponding KB identifiers, we implemented a Web tool[12] that allows a user annotate a text, producing output in the popular NLP Interchange Format (NIF) [11], as well as offering visualizations of the annotations that facilitate, e.g., revision. For each language, we provide annotated links that target the English Wikipedia entry, as well the language version of Wikipedia (if different from English). In case there was no appropriate Wikipedia entry for a mention of a person, organization or place, we annotate the mention with a `NotInLexicon` marker.

In summary, VoxEL thus consists of 15 news (documents) from the multilingual newsletter VoxEurop, totalling 94 sentences. This text is annotated times five for each language, and times two for the strict and relaxed versions, giving a total of 150 annotated documents and 940 sentences. The same number of annotations is given for each language (including by sentence). For the strict version, each language has 204 annotated mentions, while for the relaxed version, each language has 674 annotated mentions. Again, this homogeneity of text and anno-

[12] `https://users.dcc.uchile.cl/~hrosales/NIFify.html`

tations across languages was non-trivial to achieve, but facilitates comparison of evaluation results not only across systems, but across languages. For comparison, the SemEval dataset has 758, 768 and 784 annotations in the English, Spanish and Italian version respectively, while MEANTIME has 2790, 2729, 2709 and 2704 mentions in the English, Dutch, Italian and Spanish versions of their text.

5 Experiments

We now use our proposed VoxEL dataset to conduct experiments in order to explore the behavior of state of the art EL systems for multilingual settings. In particular, we are interested in the following questions:

- **RQ1**: How does the performance of systems compare for multilingual EL?
- **RQ2**: For which of the five languages are the best results achieved?
- **RQ3**: How would a method based on machine translation to English compare with directly configuring the system for a particular language?

In order to address **RQ1** and **RQ2**, we ran the multilingual EL systems Babelfy, AGDISTIS, DBpedia Spotlight, FREME, TagME and THD over both versions of VoxEL in all languages. These experiments were conducted with the GERBIL [31] EL evaluation framework, which provides unified access to the public APIs of multiple EL tools, abstracting different input and output formats using the NIF vocabulary, allowing to apply standard metrics to measure the performance of results with respect to a labelled dataset. Note that we had to slightly modify GERBIL to be able to configure the available EL systems for different languages; previously it would only support English configurations. GERBIL uses these systems via their REST APIs maintaining default (non-language) parameters, except for the case of Babelfy, for which we analyze two configurations: one that applies a more liberal interpretation of entities to include conceptual entities (Babelfy$_R$), and another configuration that applies a stricter definition of entities (Babelfy$_S$), where the two configurations correspond loosely with the relaxed/strict versions of our dataset.

The results of these experiments are shown in Table 3, where we present micro-measures for Precision (mP), Recall (mR) and F_1 (mF), with all systems, for all languages, in both versions of the dataset.[13] From first impressions, we can observe that two systems – TagME and THD – do not support all languages, where we leave the corresponding result in blank. On the other hand, the API for FREME threw exceptions when set for languages other than English.

With respect to **RQ1**, for the Relaxed version, the highest F_1 scores are obtained by Babelfy$_R$ (0.662: ES) and DBspot (0.650: EN), which have comparable performance across languages. On the other hand, the highest F_1 scores for the Strict version are TagME (0.857: EN) and Babelfy$_S$ (0.805: ES). In general,

[13] The GERBIL results are available at `https://users.dcc.uchile.cl/~hrosales/ISWC2018_experiment_GERBIL.html`

Table 3. GERBIL Evaluation of EL systems with Micro Recall, Precision and F_1. We abbreviate *DBpedia Spotlight* as *DBspot*. A value *err* indicates that an error was encountered, while "–" indicates that the system does not support the corresponding language. The results in bold are the best for that metric, system and dataset variant comparing across the five languages (the best in each row, split by Relax/Strict).

		Relaxed					Strict				
		DE	EN	ES	FR	IT	DE	EN	ES	FR	IT
Babelfy$_R$	mP	**0.840**	0.649	0.835	0.824	0.810	**0.932**	0.785	0.929	0.889	0.907
	mR	0.181	0.522	**0.549**	0.488	0.451	0.676	**0.735**	0.710	0.632	0.578
	mF	0.595	0.578	**0.662**	0.613	0.579	0.784	0.759	**0.805**	0.739	0.706
Babelfy$_S$	mP	0.903	0.722	**0.916**	0.912	0.884	**0.942**	0.816	0.923	0.912	0.894
	mR	**0.461**	0.219	0.210	0.200	0.192	0.558	0.524	**0.593**	0.563	0.583
	mF	0.301	0.336	**0.342**	0.328	0.316	0.701	0.638	**0.722**	0.697	0.706
AGDISTIS	mP	0.313	0.348	0.437	0.282	**0.512**	0.568	**0.779**	0.549	0.475	0.725
	mR	0.237	0.265	0.329	0.210	**0.397**	0.568	**0.779**	0.549	0.475	0.725
	mF	0.270	0.301	0.376	0.241	**0.447**	0.568	**0.779**	0.549	0.475	0.725
DBspot	mP	0.731	**0.745**	0.691	0.658	0.682	0.781	**0.854**	0.690	0.691	0.800
	mR	0.508	**0.577**	0.399	0.360	0.488	0.544	**0.602**	0.382	0.406	0.549
	mF	0.600	**0.650**	0.506	0.466	0.569	0.641	**0.706**	0.492	0.512	0.651
FREME	mP	*err*	**0.781**	*err*	*err*	*err*	*err*	**0.872**	*err*	*err*	*err*
	mR	*err*	**0.370**	*err*	*err*	*err*	*err*	**0.402**	*err*	*err*	*err*
	mF	*err*	**0.503**	*err*	*err*	*err*	*err*	**0.550**	*err*	*err*	*err*
TagME	mP	0.635	**0.754**	–	–	0.494	0.875	**0.946**	–	–	0.742
	mR	0.232	**0.488**	–	–	0.182	0.652	**0.784**	–	–	0.509
	mF	0.340	**0.592**	–	–	0.266	0.747	**0.857**	–	–	0.604
THD	mP	**0.831**	0.806	–	–	–	**0.857**	0.809	–	–	–
	mR	0.109	**0.253**	–	–	–	0.352	**0.647**	–	–	–
	mF	0.194	**0.386**	–	–	–	0.500	**0.719**	–	–	–

the F_1 scores for the Strict version were higher than those for the Relaxed version: investigating further, the GERBIL framework only considers annotations to be false positives when a different annotation is given in the labelled dataset for an overlapping position; hence fewer labels in the Strict dataset will imply fewer false positives overall, which seems to outweigh the effect of the additional true positives that the Relaxed version would generate. Comparing the best Strict/Relaxed results for each system, we can see that Babelfy$_R$, DBspot and FREME have less of a gap between both, meaning that they tend to annotate a broader range of entities, while, in particular, Babelfy$_S$ AGDISTIS and THD are more restrictive in the types of entities that they link.

With respect to **RQ2**, considering all systems, we can see a general trend that English had the best results overall, where it was always best for DBspot, FREME and TagME. For THD, German had higher precision, whereas English had higher recall. On the other hand, Babelfy generally had best results in German and Spanish, often having the *lowest* precision in English, while AGDISTIS seems to favour Italian for the relaxed dataset. There are variations between languages that may make the EL task easier or harder depending on the features used; for example, systems that rely on capitalization may perform differently

for Spanish, which uses less capitalization, (e.g., "*Jungla de cristal*": a Spanish movie title in sentence case); and German, where all nouns are capitalized.

Regarding **RQ3**, we will require another experiment to address the question of the efficacy of using machine translations. First we note that, although works in related areas – such as cross-lingual ontology matching [9] – have used machine translation to adapt to multilingual settings, to the best of our knowledge, no system listed in Table 1 uses machine translations over the input text (though systems such as Babelfy do use machine translations to enrich the lexical knowledge available in the KB). Hence we check to see if translating a text to English using a state-of-the-art approach – Google translator[14] – and applying EL over the translated English text would fare better than applying EL directly over the target language; we choose one target language to avoid generating results for a quadratic pairing of languages, and we choose English since it was the only language working for/supported by all systems in Table 3.

A complication for these translation experiments is that while VoxEL contains annotations for the texts in their original five languages, including English, it does not contain annotations for the texts translated to English. While we considered manually annotating such documents produced by Google Translate, we opted against it partly due to the amount of labour it would again involve, but more importantly because it would be specific to one translation service at one point in time: as these translation services improve, these labelled documents would quickly become obsolete. Instead, we apply evaluation on a per-sentence basis, where for each sentence of a text in a non-English language, we translate it and then compare the set of annotations produced against the set of manually-annotated labels from the original English documents; in other words, we check the annotations produced by sentence, rather than by their exact position. This is only possible because in the original VoxEL dataset, we defined a one-to-one correspondence between sentences across languages.

Note that since GERBIL requires labels to have a corresponding position, we thus needed to run these experiments locally outside of the GERBIL framework. Hence, for a sentence s, let A denote the IRIs associated with manual labels for s in the original English text, and let B denote the IRIs annotated by the system for the corresponding sentence of the translated text; we denote true positives by $A \cap B$, false positives by $B - A$, and false negatives by $A - B$.[15]

In Table 4, we show the results of this second experiment, focusing this time on the Micro-F_1 score obtained for each system over the five languages of VoxEL, again for the relaxed and strict versions. For each system, we consider three experiments: (1) the system is configured for the given language and run over text for the given language, (2) the system is configured for English and run over the text translated from the given language, (3) the system is configured for English and run over the text in the given language without translation. We use the third configuration to establish how the translation to English – rather than the system configuration to English – affects the results. First we

[14] `https://translate.google.com/`; April 1st, 2018

[15] To compute Precision, Recall and F_1, we do not require true negatives.

Table 4. Micro F_1 scores for systems performing EL with respect to the VoxEL dataset. For each system and each non-English language, we show the results of three experiments: first, for (_,_) the system is configured for the same language as the current text; second, for (EN,EN$_t$), the system is configured for English and applied to text translated to English from the original language; third, for (EN,_), the system is configured for English and run for the text in the current (original) language. Below the name of each system, we provide the relaxed and strict results for English text. Underlined results indicate the best of the three configurations for the given system, language and dataset variant (e.g., best for the columns of three values). Bold results indicate the best result for that system across all variations.

		Relaxed				Strict			
		DE	ES	FR	IT	DE	ES	FR	IT
Babelfy$_R$ (0.545,0.319)	(_,_)	0.523	**0.541**	0.493	0.504	0.344	0.362	0.309	0.365
	(EN,EN$_t$)	0.507	0.515	0.505	0.501	0.298	0.298	0.314	0.301
	(EN,_)	0.215	0.170	0.195	0.140	0.253	0.239	0.220	0.179
Babelfy$_S$ (0.308,0.567)	(_,_)	0.279	0.325	0.290	0.311	0.572	0.611	0.583	**0.616**
	(EN,EN$_t$)	0.311	0.309	0.322	0.303	0.518	0.523	0.559	0.532
	(EN,_)	0.201	0.179	0.189	0.137	0.376	0.372	0.395	0.258
DBspot (0.466,0.707)	(_,_)	0.400	0.331	0.240	0.342	0.510	0.477	0.481	0.653
	(EN,EN$_t$)	0.441	0.454	0.464	0.449	0.696	0.694	0.721	**0.729**
	(EN,_)	0.209	0.161	0.18	0.188	0.374	0.259	0.326	0.323
FREME (0.407,0.708)	(_,_)	0.282	0.302	0.268	0.373	0.483	0.583	0.479	**0.726**
	(EN,EN$_t$)	0.404	0.403	0.401	0.408	0.701	0.713	0.692	0.711
	(EN,_)	0.166	0.183	0.196	0.222	0.190	0.338	0.342	0.374
TagME (0.462,0.327)	(_,_)	0.414	–	–	-	0.272	–	–	–
	(EN,EN$_t$)	0.431	**0.450**	0.441	0.439	0.330	0.333	0.321	0.336
	(EN,_)	0.188	0.181	0.200	0.148	0.212	0.202	0.197	0.164
THD (0.392,0.625)	(_,_)	0.241	–	–	–	0.546	–	–	–
	(EN,EN$_t$)	0.394	0.392	0.386	0.387	0.597	0.620	0.595	**0.623**
	(EN,_)	0.207	0.175	0.217	0.174	0.251	0.332	0.403	0.352

note that without using positional information to check false positives (as per GERBIL), the results change from those presented in Table 3; more generally, the gap between the Relaxed and Strict version is reduced.

With respect to RQ3, in Table 4, for each system, language and dataset variant, we underline which of the three configurations performs best. For example, in DBspot, all values on the (EN,EN$_t$) line – which denotes applying DBspot configured for English over text translated to English – are underlined, meaning that for all languages, prior translation to English outperformed submitted the text in its original language to DBspot configured for that language.[16] In fact, for almost all systems, translating the input text to English generally outperforms using the available language configurations of the respective EL systems, with the exception of Babelfy, where the available multilingual settings generally outperforms the prior translation to English (we may recall that in Table 3,

[16] ... it also implies that it outperforms running English EL on text in the original language, though this is hardly surprising and just presented for reference.

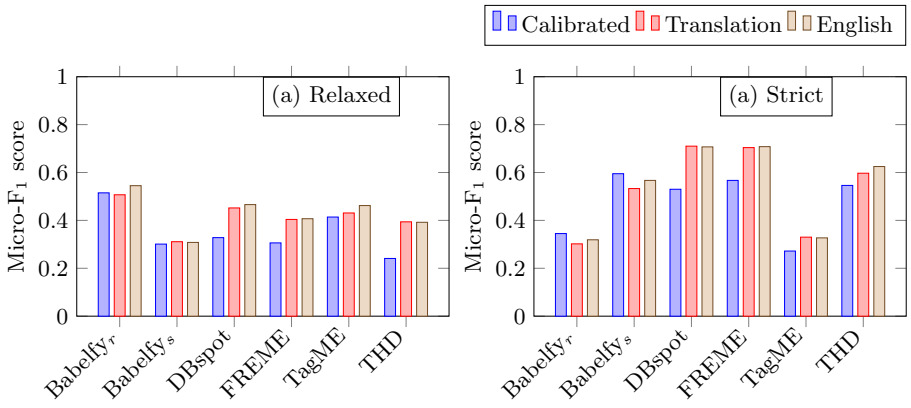

Fig. 1. Summary of the MicroF$_1$ results over VoxEL Relaxed/Strict for the translation experiments, comparing mean values for setting the EL system to the language of the text (*Calibrated*), translating the text to English first (*Translation*), and the corresponding F_1 score for EL over the original English text (*English*)

Babelfy performed best in texts other than English). We further note that the translation results are generally competitive with those for the original English text – shown below the name of the system for the Relaxed and Strict datasets – even slightly outperforming those results in some cases. We can further observe from the generally poor (EN,_) results that the translation is important; in other words, one cannot simply just apply an EL system configured for English over another language and expect comparable results.

To give a better impression of the results obtained from the second experiment, in Figure 1, for the selected systems, we show the following aggregations: (1) *Calibrated*: the mean Micro-F_1 score across the four non-English languages with the EL system configured for that language; (2) *Translation*: the mean Micro-F_1 score across the four non-English languages with the text translated to English and the EL system configured for English; (3) English, which presents the (single) Micro-F_1 score for the original English text. From this figure, we can see visually that translation is comparable to native English EL, and that translation often considerably outperforms EL in the original language.

6 Conclusion

While traditionally Entity Linking has mostly focussed on processing texts in English, in recent years there has been a growing trend towards developing techniques and systems that can support multiple languages. To support such research, in this paper we have described a new labelled dataset for multilingual EL, which we call VoxEL. The dataset contains 15 new articles in 5 different languages with 2 different criteria for labelling, resulting in a corpus of 150 manually-annotated news articles. In a Strict version of the dataset considering

a core of entities, we derive 204 annotated mentions in each languages, while in a Relaxed version of the dataset considering a broader range of entities described by Wikipedia, we derive 674 annotated mentions in each language. The VoxEL dataset is distinguished by being based on texts translated by experts, having a one-to-one correspondence of sentences – and annotated entities per sentence – between languages. The dataset (in NIF) is available online under a CC-BY 4.0 licence: `https://dx.doi.org/10.6084/m9.figshare.6104759`.

We then use the VoxEL dataset to conduct experiments comparing the performance of selected EL systems in a multilingual setting. We found that in general, Babelfy and DBpedia Spotlight performed the most consistently across language. We also found that with the exception of Babelfy, the results for EL systems were best over the English version of the texts. Next, we compared configuring the multilingual EL system for each non-English language versus applying a machine translation of the text to English and running the system in English; with the exception of Babelfy, we found that the machine translation approach outperformed configuring the system for a non-English language; even in the case of Babelfy, in some cases the translation performed better, while in general both were deemed comparable. This raises an important issue for research on multilingual EL: state-of-the-art machine translation is now reaching a point where we must ask if it is worth building dedicated multilingual EL systems, or if we should focus on EL for one language to which other languages can be machine translated.

Acknowledgements The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also supported by the Millennium Nucleus Center for Semantic Web Research under Grant NC120004. We would like to thank Michael Röder for his considerable help regarding GERBIL.

References

1. Brando, C., Frontini, F., Ganascia, J.G.: Reden: named entity linking in digital literary editions using linked data sets. CSIMQ (7), 60–80 (2016)
2. Brümmer, M., Dojchinovski, M., Hellmann, S.: Dbpedia abstracts: A large-scale, open, multilingual nlp training corpus. In: LREC (2016)
3. Charton, E., Gagnon, M., Ozell, B.: Automatic semantic web annotation of named entities. In: Canadian AI. pp. 74–85. Springer (2011)
4. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. EMNLP-CoNLL p. 708 (2007)
5. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: I-SEMANTICS. pp. 121–124 (2013)
6. Dojchinovski, M., Kliegr, T.: Recognizing, classifying and linking entities with Wikipedia and DBpedia. WIKT pp. 41–44 (2012)
7. Fahrni, A., Göckel, T., Strube, M.: HITS' monolingual and cross-lingual entity linking system at TAC 2012: A joint approach. In: TAC. Citeseer (2012)
8. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM. pp. 1625–1628. ACM (2010)

9. Fu, B., Brennan, R., O'sullivan, D.: Cross-lingual ontology mapping and its use on the multilingual semantic web. MSW 571, 13–20 (2010)
10. Guo, Z., Xu, Y., de Sá Mesquita, F., Barbosa, D., Kondrak, G.: ualberta at TAC-KBP 2012: English and cross-lingual entity linking. In: TAC (2012)
11. Hellmann, S., et al.: Integrating nlp using linked data. In: ISWC. pp. 98–113 (2013)
12. Hoffart, J., et al.: Robust disambiguation of named entities in text. In: EMNLP. pp. 782–792. ACL (2011)
13. Hoffart, J., Seufert, S., Nguyen, D.B., Theobald, M., Weikum, G.: Kore: keyphrase overlap relatedness for entity disambiguation. In: CIKM. pp. 545–554 (2012)
14. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: SIGKDD. pp. 457–466 (2009)
15. Lehmann, J.e.a.: DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2), 167–195 (2015)
16. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: I-SEMANTICS. pp. 1–8. ACM (2011)
17. Minard, A., et al.: Meantime, the newsreader multilingual event and time corpus (2016)
18. Moro, A., Navigli, R.: SemEval-2015 Task 13: Multilingual all-words sense disambiguation and entity linking. In: SemEval@ NAACL-HLT. pp. 288–297 (2015)
19. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Trans. of the ACL 2, 231–244 (2014)
20. Moussallem, D., et al.: Mag: A multilingual, knowledge-base agnostic and deterministic entity linking approach. In: K-CAP. p. 9 (2017)
21. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: WSDM. pp. 365–374. ACM (2017)
22. Popov, B., et al.: Kim–a semantic platform for information extraction and retrieval. Nat. Lang. Eng. 10(3-4), 375–392 (2004)
23. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to Wikipedia. In: NAACL-HLT. pp. 1375–1384 (2011)
24. Röder, M., et al.: N^3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In: LREC. pp. 3529–3533 (2014)
25. Rosales-Méndez, H., et al.: What should entity linking link? In: AMW (2018)
26. Rosales-Méndez, H., Poblete, B., Hogan, A.: Multilingual entity linking: Comparing english and spanish. In: LD4IE@ISWC. pp. 62–73 (2017)
27. Sasaki, F., Dojchinovski, M., Nehring, J.: Chainable and extendable knowledge integration web services. In: ISWC. pp. 89–101 (2016)
28. Speck, R., et al.: Ensemble learning of named entity recognition algorithms using multilayer perceptron for the multilingual web of data. In: K-CAP. p. 26 (2017)
29. Tsai, C.T., Roth, D.: Cross-lingual wikification using multilingual embeddings. In: NAACL-HLT. pp. 589–598 (2016)
30. Usbeck, R., et al.: AGDISTIS-graph-based disambiguation of named entities using linked data. In: ISWC. pp. 457–471. Springer (2014)
31. Usbeck, R., et al.: GERBIL: general entity annotator benchmarking framework. In: WWW. pp. 1133–1143 (2015)
32. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
33. Waitelonis, J., Exeler, C., Sack, H.: Linked data enabled generalized vector space model to improve document retrieval. In: NLP & DBpedia @ ISWC (2015)
34. Wang, Z., Li, J., Tang, J.: Boosting cross-lingual knowledge linking via concept annotation. In: IJCAI. pp. 2733–2739 (2013)