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. We then present the results of experiments using this dataset to compare the behavior of state of the art of (multilingual) EL systems across the five different languages, additionally comparing these results with methods using machine translation to English.

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 semantic information of publicly available KBs about real-world entities to achieve a better understanding of natural language text. 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 [19] (`wd:Q2831`), or the DBpedia KB [8] (`dbr:Michael_Jackson`)[1] and thereafter leverage the knowledge in the KB about the entity mentioned in the text to support semantic search, relationship extraction, 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

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

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 besides – 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 [9,7]. These approaches often use resources of a specific language, such as Part-Of-Speech taggers and WordNet[2], that 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 [**?**], DBpedia Spotlight Corpus[9], KORE 50 [**?**]).

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 [3,2,5,**?**] – relying only on the language of labels available in the reference KB – or can be configured for multiple languages [11,**?**]. Recognising 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 [14] 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 performance across languages. 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, English, Spanish, French and Italian languages. This dataset is based on an online source of multilingual news, where we select and annotate 15 corresponding news articles for these five languages (giving 45 articles in total). Additionally we create 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

[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.

on traditional MUC/NER definitions), and a *relaxed* version where we consider a broader selection of mentions referring to entities described by Wikipedia. Based on the VoxEL dataset, using the GERBIL evaluation framework [18], 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.

Outline: Section 2 introduces preliminaries relating to the Entity Linking task. Section 3 discusses works that have address the problem of multilingual Entity Linking, including discussion of both systems and datasets. Section 4 presents the design and creation of the VoxEL dataset, along with some descriptive statistics. Section 5 presents some research questions that can be addressed with VoxEL, and presents corresponding experiments over a selection of EL systems. We then conclude and discuss future directions in Section 6.

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 such mention is typically a string considered as referring 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:

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 [9,11]. However, other approaches may instead apply a unified process, building models that create feedback between the recognition and disambiguation steps [**?**]. 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 [14], even a system supporting only English components may still be able to correctly recognise 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, where for example language-agnostic systems – systems that require no linguistic components or resources specific to a language, typically ones based on a dictionary approach – 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 summarise 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 (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 [13] – most of the included multilingual EL approaches have emerged since 2011.

[6] We presented an earlier version of such a table in previous work [14].

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 [13]	2004	English, French, Spanish	✓	✗	✓
SDA [1]	2011	English, French	✗	✗	✗
ualberta [6]	2012	English, Chinese	✗	✗	✗
HITS [4]	2012	English, Spanish, Chinese	✗	✗	✗
THD [3]	2012	English, German, Dutch	✓	✓	✓
DBpedia Spotlight [9,2]	2013	English, Italian, Russian, Dutch, French, German, Spanish, Hungarian, Danish	✓	✓	✓
TagME [5]	2013	English, German, Dutch	✓	✗	✓
Wang-Tang [20]	2013	English, Chinese	✗	✗	✗
AGDISTIS [16,**?**]	2014	English, German, Spanish French, Italian, Japanese, Dutch	✓	✓	✓
Babelfy [11]	2014	English, Spanish, French German, Italian	✓	✗	✓
FREME [**?**]	2016	English, German	✗	✓	✓
WikiME [15]	2016	English, Spanish, French, Italian, Chinese, German, Thai, Arabic, Turkish, Tamil, Tagalog, Urdu, Hebrew	✓	✗	✗
FEL [12]	2017	English, Spanish, Chinese	✗	✓	✗
FOX [**?**]	2017	English, German, Spanish, French, Dutch	✓	✓	✓

We will later conduct experiments using the GERBIL evaluation framework [17], 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 [3]: *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 [9], 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. [2] 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 [5]. 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 [16]. 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 [**?**] 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 [11]. 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 [**?**].

AH: Any explanation for not including FOX?

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

[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 [**?**]	1393	22,137	34,929	✓	EN
AIDA/CoNLL-Test A [**?**]	216	3,466	5,917	✓	EN
AIDA/CoNLL-Test B [**?**]	231	3,684	5,616	✓	EN
AIDA/CoNLL-Training [**?**]	946	14,987	23,396	✓	EN
KORE50 [**?**]	50	50	144	✓	EN
IITB [**?**]	103	1,781	18,308	✓	EN
ACE2004 [**?**]	57	-	306	✗	EN
AQUAINT [**?**]	50	533	727	✗	EN
MSNBC [**?**]	20	668	747	✗	EN
DBpedia Spotlight [9]	10	58	331	✓	EN
N3-RSS 500 [**?**]	1	500	1000	✓	EN
Reuters 128 [**?**]	128	-	881	✓	EN
Wes2015 [**?**]	331	-	28,586	✓	EN
News-100 [**?**]	100	-	1656	✓	DE
Thibaudet [**?**]	1	3,807	2,980	✗	FR
Bergson [**?**]	1	4,280	380	✗	FR
SemEval 2015 Task 13 [10]	4	137	769	✓	EN,ES,IT
DBpedia Abstracts [**?**]	39,132	-	505,033	✗	DE,EN,ES,FR, IT,JA,NL
MEANTIME [**?**]	120	597	2,790	✗	EN,ES,IT,NL
VoxEL$_R$	15	94	674	✓	DE,EN,ES,IT,FR
VoxEL$_S$	15	94	204	✓	DE,EN,ES,IT,FR

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).

From a brief inspection, we can see that the majority of datasets provide text in English only, 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 [10].

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 [**?**].[8]

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

Amongst these multilingual datasets, we can highlight a number of shortcomings. First note that we do not include the 2013 TAC–KBP dataset in this list since it was not available at the time of writing.

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 [5] and FREME [**?**] 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 [14] 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 (similiar to the traditional NER/TAC consensus on an entity) AH: refs from AMW paper, 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 AH: refs from AMW paper. Furthermore, there is a lack of consensus on how overlapping entities – like *New York City Fire Department* – should be treated AH: ref from AMW; 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.

AH: Henry: could you add an example strict and relaxed sentence in English for comparison? Something relatively clean?

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) [**?**], 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 conduct some experiment in order to explore the behavior of state of the art EL systems in both version of the VoxEL dataset. In particular, we are interested in the following questions:

- **RQ1**: How does EL performance differ between the languages German, English, Spanish, French and Italian using VoxEL?
- **RQ2**: Could automatic translation be used to achieve results similar to that achieved in the English language by monolingual systems applied to texts written in a language that does not have its configuration?
- **RQ3**: Taking into account that the systems obtain a better performance for English, could the automatic translation overcome the results of such systems for the other languages that support VoxEL?.

To response RQ1 we ran the multilingual systems Babelfy, AGDISTIS, DBpedia Spotlight, FREME NER, TagME and THD over both version of VoxEL. The results are shown in Table 3. We use these systems via their REST APIs maintaining its predetermined parameters, except for the case of Babelfy for which we analyze separately the inclusion (Babelfy$_r$) and not inclusion (Bablfy$_s$) of concepts in the annotations. This first experiment was conducted by designing a modified local version of GERBIL [**?**], which is a friendly platform with a suits of annotator and datasets available for agile comparisons. So far, GERBIL only consider English comparison, therefore, we include missing configurations in order that multilingual systems could annotate non-English texts.

We show in Table 3 the Micro F_1 obtained in the experiment[13]. While DBpedia Spotlight and FREME NER show a similar behavior in both environment, Babelfy, AGDISTIS, TagME and THD are sensitives to the inclusion of entities that do not are persons, organizations or places. In all of the cases, the involved systems agree with the fact that the strict version of VoxEL better fits the task needs.

Incorporating more than one language into an EL approach brings with it an increase in its complexity. For instance, WikiME uses a model based on word embedding, which includes a final step for each foreign language embeddings of projection to the English one. On the other hand, Babelfy exploits the sets of synonyms of BabelNet[14], called *Babel synsets*, that groups words in different

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

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

Table 3. Evaluation of EL systems according GERBIL using Micro F_1. We refer FREME NER and DBpedia Spotlight as *FREME* and *DBspot* respectively. In the cases when no result was obtained we show *err*, as well as "-" when a system does not support the corresponding language.

		Relax					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	-	-	-

languages to achieve multilingual EL. According to our review, no system has used machine translation to address the language barriers in EL. Some work with automatic translation is done in related areas, such as cross-lingual ontology matching [**?**] where the impact of the translation quality is stressed.

In this direction, we design a second experiment in order to study the behavior of these same approaches configured only for English over the non-English version of VoxEL by automatic translation to English. For the automatic translation we use Google translator[15]. VoxEL contains links to the language for each of its versions and also to English, what makes it possible to evaluate automatic translations directly. However, when we apply the translation we lost the details of the annotation information that it is required by GERBIL, such as the translated mention, initial and final position of the mention, etc. For this reason, we implement our own benchmark that leaves out the mentions information of the evaluation process, only considering the match between the URIs of the gold standard and the system results.

In Table 4 we show the results of the second experiment, showing the MicroF_1 score obtained for each system over the version of VoxEL written in its five languages. We configure the system with English and ran then over the version

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

Table 4. Performance of some EL systems using automatic translation according to F_1. For each system Sys and the current language L we show three experiments, first (_,_) denotes the behavior of Sys configured with L over the dataset wrote in this same language. The second (EN,EN$_t$) shows the results that involve the system Sys configured with English and applied over an automatic translation of the datasets written in the current language to English. The third, denoted by (EN,_), corresponds to the evaluation of Sys configured with English and applied over the datasets written in the current language L. The information below to each system name correspond to (EN,EN) evaluation for the relax and strict version of VoxEL, in this order.

		Relax				Strict			
		DE	ES	FR	IT	DE	ES	FR	IT
Freme NER (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
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
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
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

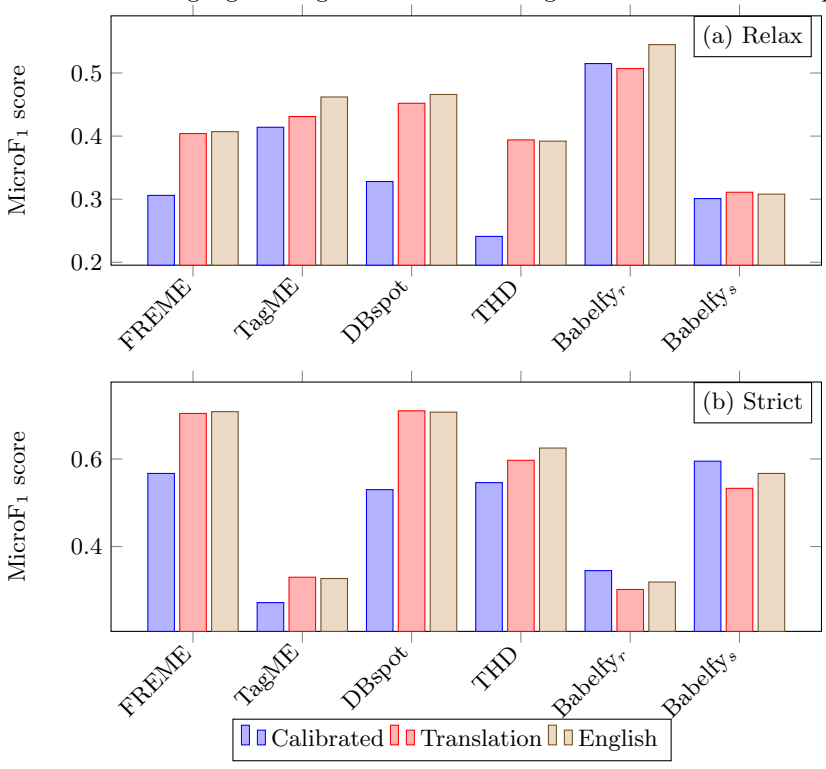
of VoxEL in the original language L and also in the translated version, with the aim of having a point of reference that allows to mediate a worse and better behavior of the system on VoxEL. We took the first of both performances as *base line* for each system and the second one as *desirable performance*. As we can observe in Table 4, the score from EL applied for the translation is between both criteria in all the cases, which indicates that the incorporation of automatic translation in this context obtains reasonable results. Thus, we respond the RQ2.

We underlined on in Table 4 for each system those scores that represent a better performance through automatic translation than the baseline and the desirable performance. As we can observe, this happens for all the systems, at least for one of the languages. In particular for DBpedia Spotlight, TagME and THD this always happens. On the other hand, we also highlight with bold font those scores that represent for each system the better performance among all its scores. This behavior is presented by three of the six systems considered in the experiment. So, this response RQ3, placing automatic translation as an

option to tackle multilingual EL from systems with missing languages in their configuration as is the case of TagME and THD in this experiment.

To give a better idea of the results obtained from the second experiment, we show in Figure 1 the average score of the selected systems in each of the languages supported by VoxEL. In both version of VoxEL we can observe that in average, the automatic translation is a way of approaching multilingual EL comparable with models that become more complex when they support more than one language.

Fig. 1. Average of the MicroF$_1$ evaluation that has the systems selected in each language supported by VoxEL, over its both version: (a) relaxed and the (b)strict. We show the *Calibrated* values corresponding to the performance of each system configured by the current language, over the versions of VoxEL written in this same language. On the other hand, the *Translation* and *English* values are relating to the performance of each language configured for English and applied over the translation of version written in the current language to English and over the English version of VoxEL respectively.



6 Conclusion

There are several multilingual approaches to address EL in the literature, as well as benchmark datasets that serve as a gold standard in the evaluation process. In this work we propose a manually annotated VoxEL, a new dataset that meets a set of properties that we consider desirable in multilingual EL assessments. Among these properties, we have that VoxEL has notes on cured texts, and not on texts translated automatically as is the case of some multilingual datasets of literature. This benefits to a greater extent to evaluate in a fair way the system that bases its models on intrinsic aspects of the lexicon. Also, we advocate for multilingual datasets that contain the same number of documents, sentences and annotations, in order to provide results that allow us the intra-language behavior of the systems. Property fulfilled by VoxEL.

We then conduct experiments to compare the performance of selected multilingual EL systems over VoxEL. We found an expected behavior in systems, being Babelfy and DBpedia Spotlight the best scored. In general, all the systems obtained better quality in the results for the annotations that only target persons, organizations and places. We also conducted experiments using VoxEL, to study how these systems perform when they are configured for the English language and an automatic translation is used to achieve multilingual annotation.

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 response some helpful email about GERBIL functionalities.

References

1. Charton, E., Gagnon, M., Ozell, B.: Automatic semantic web annotation of named entities. In: Canadian Conference on Artificial Intelligence. pp. 74–85. Springer (2011)
2. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: I-SEMANTICS. pp. 121–124. ACM (2013)
3. Dojchinovski, M., Kliegr, T.: Recognizing, classifying and linking entities with Wikipedia and DBpedia. WIKT pp. 41–44 (2012)
4. Fahrni, A., Göckel, T., Strube, M.: HITS' monolingual and cross-lingual entity linking system at TAC 2012: A joint approach. In: TAC. Citeseer (2012)
5. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM. pp. 1625–1628. ACM (2010)
6. 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)
7. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP. pp. 782–792. ACL (2011)

8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2), 167–195 (2015)
9. 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)
10. Moro, A., Navigli, R.: SemEval-2015 Task 13: Multilingual all-words sense disambiguation and entity linking. In: SemEval@ NAACL-HLT. pp. 288–297 (2015)
11. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Trans. of the ACL 2, 231–244 (2014)
12. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: WSDM. pp. 365–374. ACM (2017)
13. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: Kim–a semantic platform for information extraction and retrieval. Natural Language Engineering 10(3-4), 375–392 (2004)
14. Rosales-Méndez, H., Poblete, B., Hogan, A.: Multilingual entity linking: Comparing english and spanish. In: International Workshop on Linked Data for Information Extraction (LD4IE). pp. 62–73 (2017)
15. Tsai, C.T., Roth, D.: Cross-lingual wikification using multilingual embeddings. In: NAACL-HLT. pp. 589–598 (2016)
16. Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: AGDISTIS-graph-based disambiguation of named entities using linked data. In: ISWC. pp. 457–471. Springer (2014)
17. Usbeck, R., Röder, M., Ngomo, A.C.N.: Evaluating entity annotators using GERBIL. In: ESWC (Satellite Events). pp. 159–164 (2015)
18. Usbeck, R., Röder, M., Ngomo, A.N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., Wesemann, L.: GERBIL: general entity annotator benchmarking framework. In: WWW. pp. 1133–1143 (2015)
19. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
20. Wang, Z., Li, J., Tang, J.: Boosting cross-lingual knowledge linking via concept annotation. In: IJCAI. pp. 2733–2739 (2013)