Towards Better Entity Linking Evaluation

Henry Rosales-Méndez
DCC, University of Chile
Supervised by: Aidan Hogan and Barbara Poblete

ABSTRACT

The Entity Linking (EL) task is concerned with linking entity mentions in a text collection with their corresponding knowledge-base entries. Despite the progress made in the evaluation of EL systems, there is still much work to be done, where this Ph.D. research tackles issues concerning EL evaluation. Among these issues, we stress (a) the lack of consensus about the definition of "entity" and the lack of evaluation metrics that allow for different notions of entities, (b) the lack of datasets that allow for cross-language comparison, and (c) the focus on evaluating high-level systems rather than low-level techniques. By addressing these challenges and better understanding the performance of EL systems, our hypothesis is that we can create a more general, more configurable EL framework that can be better adapted to the needs of a particular application. In the early stages of this PhD work, we have identified these problems and begun to address (a) and (b), publishing initial results that constitute a significant step forward in our investigation. However, there are still further challenges that must be addressed before we reach our goal. Our next steps thus involve proposing a more fluid definition of "entity" adaptable to different applications, the definition of quality measures that allow for comparing EL approaches targeting different types of entities, as well as the creation of a customizable EL framework that allows for composing and evaluating individual techniques as appropriate to a particular task.

CCS CONCEPTS

• **Information systems** → *Information extraction.*

KEYWORDS

Information Extraction; Entity Linking; Quality measurement

ACM Reference Format:
Henry Rosales-Méndez. 2019. Towards Better Entity Linking Evaluation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308558.XXXXXXX

1 PROBLEM

Entity Linking (EL) is a task in Information Extraction (IE) that focuses on linking the entity mentions in a text collection with entity identifiers in a given Knowledge Base (KB). Such a task has various applications, including semantic search, document classification, semantic annotation, and text enrichment, as well as forming the basis for further IE processes. Despite the fact that works addressing the EL tasks have been pursued by various communities and published in various international conferences, some fundamental questions remain open regarding the aim of the task and how EL results should be evaluated.

A [second]^{t}, larger and more [theatrical]^{t} [Cirque]^{t} [show]^{t}, [{Michael Jackson}^{btdf}: One]^{bt}, designed for [residency]^{t} at the [Mandalay Bay]^{btdf} [resort]^{d} in [Las {Vegas}^{b}]^{td}.

Figure 1: Output annotations of Babelfy (b), TagME (t), DBpedia Spotlight (d) and FRED (f) over the same input

First and foremost, despite the presence of various gold standard datasets, evaluation frameworks, etc., it is still unclear what EL systems should link. There is evidence of disagreement in the EL community on this matter, and as a consequence, different systems target different types of entities. This phenomenon is illustrated in Figure 1, which contains the results for a short example text of four state-of-the-art systems that are popular in the community: Babelfy [23], DBpedia Spotlight [18], TagMe [7] and FRED [9]. As we can observe in Figure 1, there are signs of fundamental disagreements among the involved systems. While FRED and Babelfy consider only proper names, TagME and DBpedia Spotlight also include other nouns for which a corresponding KB entity exists and which do not constitute names. Furthermore, overlapping mentions (denoted by "{}") are targeted by Babelfy and TagME, but not by DBpedia Spotlight nor FRED. So which system is more correct?

This lack of consensus affects further processing of EL systems' outputs since different application scenarios have different requirements on what mentions should be involved. Furthermore, this problem also complicates EL assessment because we do not know how we can define the ideal result that such a system should achieve. Some efforts have been made to standardize which mentions we should identify for annotation, as is the case of the work by Jha et al. [15], who propose a set of rules to serve as guidelines for benchmark creation. However, these rules force the adoption of some considerations that may not suit certain applications and on which there is thus no consensus. For instance, Jha et al., advocate for the omission of overlapping mentions like "{Michael Jackson}", but authors such as Ling et al. [17] disagree. In a semantic search scenario, for example, looking at Figure 1, should such a document be considered relevant for a user interested in texts about Michael Jackson, or more generally, texts about American pop singers?

Several EL benchmark datasets have been proposed that – although used by a variety of systems – also exhibit this disagreement. While KORE50 [12] only annotates proper names, the DBpedia Spotlight dataset [18] includes annotations of common nouns such as *software* and *owner*. Additionally, the DBpedia Spotlight

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.XXXXXXX

dataset includes overlaps, for example, "Google car" is linked with `dbr:Google_self-driving_car`[1], while 'car' is linked to `dbr:Car`.

The solution thus far to address the lack of a consensus has been to define a new consensus, but we propose that a different approach is needed: that no one size fits all in terms of EL. This should not only be reflected in EL systems, but also in EL datasets and metrics used to evaluate EL systems with respect to such datasets. This presents a major challenge tackled in the context of this PhD work.

Such disagreement on what EL systems should link is not the only issue we have encountered in terms of evaluating the EL task. Another major issue is that despite some recent developments for other languages, most work has focused on English texts, both in terms of EL systems and EL datasets. Focusing on a multilingual context, some authors have proposed approaches with a large list of languages. For instance, this is the case of MAG [24], a multilingual EL system that supports annotations over 40 languages. However, the presence of such systems raises new challenges for evaluating the EL task. Such questions now include: How well do EL systems perform outside of English as a primary language? Do multilingual approaches behave equally for all of their supported languages? If not, why not? Are multilingual EL approaches really necessary with recent improvements in machine translation? These questions are not deeply studied yet in the literature. Indeed, only a few of the current EL datasets are multilingual, which complicate any kind of multilingual experimentation.

Generalizing these issues, different applications for EL may bring with them different requirements, which may be best addressed using different techniques. Aside from the issue of the types of entities and the languages targeted, there are also issues regarding for example the length of the text(s), the noise present, the domain of a text, the need to perform EL on semi-structured inputs (e.g., HTML), etc. Different EL systems proposed in the literature have been proposed to solve individual tasks. But individual systems may package together a specific set of techniques, where evaluation is conducted at the level of systems (or ensembles of systems) without understanding which techniques work best in which scenarios. Our ultimate goal, then, is to develop a EL framework that allows for composing and evaluating individual EL techniques, allowing to find the configuration best suited to a particular setting.

2 BACKGROUND

Entity Linking is a task in Information Extraction that focuses on linking the entity mentions in a text collection with entity identifiers in a given knowledge base. Formally, let E be a set of entities in a KB and M the set of entity mentions in a given text collection. The EL process focuses on linking each entity mention $m \in M$ in a text collection with an entity identifier $e \in E$ in a given Knowledge Base (KB). Generally speaking, EL models are commonly separated into two main phases, detailed below:

Entity Recognition (ER) This phase spots which phrases of the input text should be taken as mentions. This problem is also addressed by the Named Entity Recognition (NER) task, where a variety of techniques have been employed to this goal. On the other hand, some works regard ER itself as an independent task, out of the scope of EL [26].

Entity Disambiguation (ED) This phase decides which KB entities should be associated with the identified mentions. This phase is commonly divided into the following steps:

Candidate entity generation: For each entity mention $m \in M$ this stage selects E_m: a candidate set $E_m \subseteq E$ that represents entities with a high probability of corresponding to m is selected. Often this selection is based on matching m with entity labels for E in the knowledge base.

Candidate entity ranking: Each entity $e_m \in E_m$ is ranked according to an estimated confidence that it is the referent of the textual mention m. This can be performed considering a variety of features, such as the perceived "popularity" of e_m, its relation to candidates for nearby mentions, and so forth. The candidate in E_m with the best ranking may be selected as the link for m, possibly assuming it meets a certain threshold confidence (or other criteria).

Unlinkable mention prediction: Some tools consider unlinkable mentions, where no entity in the knowledge base meets the required confidence for a match to a given entity mention m. Depending on the application scenario, these mentions may be simply ignored, or may be proposed as "emerging entities" – annotated as NIL – that could be added to the knowledge base in the future.

In some more recent EL systems, the division between the EL and ED phases is less clear. Some systems apply an End-to-End approach, while other systems apply ER and EL jointly in the same model in the goal of optimizing for both tasks in one process [25, 37]. Other systems assume that entity mentions have already been identified by an existing ER approach and specifically address ED [26].

Several EL approaches have been proposed in the literature. Some of them take annotations, KB entities and their relationship as a graph and perform heuristic to find the proper matching. For instance, Babelfy [23] and AIDA [14] search the densest sub-graph applying a Random Walk with Restart and a greedy algorithm respectively. Other approaches as TagME [7] , THD [4], DBpedia Spotlight [18] and FREME [33] are based on similarity functions between mentions and the KB content. For instance, TagME ranks the candidate entities by two functions: *commonness* and *relatedness*, the first count how frequently an anchor text is linked to a particular Wikipedia entity and the second, returns how often candidate entities for different mentions are annotated from the same Wikipedia page. On the other hand, WIKIME [34] uses multilingual embedding which is trained for words and Wikipedia titles.

3 PROPOSED APPROACH/RESULTS

The present Ph.D. work proposes to address a variety of open questions regarding the evaluation of EL systems. In this direction, the following research questions are being or will be addressed:

(1) What should Entity Linking link?
   (a) How can we define the goal of the EL task?
   (b) Is consensus possible on the definition of an "entity"?
   (c) If not, how can we define benchmark EL datasets and what metrics can we use to reflect the lack of consensus?
(2) How well do EL systems perform in multilingual settings?
   (a) How can we compare EL performance across langauges?
   (b) How do EL systems perform for different languages?

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

(c) Why do results differ across languages?
(d) Are multilingual EL approaches necessary with recent improvement in machine translation?

(3) How can we adapt and configure EL techniques for different applications?
(a) What are the specific applications for EL?
(b) What are the different settings that can be considered?
(c) Which techniques work best under what assumptions?
(d) How can existing EL techniques be best configured to meet the needs of a particular application setting?

We now discuss these three high-level research questions in more detail, describing the issues faced, as well as the ongoing work and plans for future work to address them.

3.1 Lack of consensus

The concept of "named entity" was first coined by the 6th Message Understanding Conference [10] (MUC-6) where entities are assigned to one of the classes *Person*, *Location*, *Organization* and other numeric/temporal expressions. Hence the definition of an entity follows from these classes: any instance of such a class is considered an entity, and per the consensus of MUC-6, the goal of NER is to identify mentions of entities of one of these classes.

With the advent of large-scale, diverse KBs, interest grew in the EL task, which not only identifies (and types) entities in a text, but also links them to the KB. Many ER models and ER benchmark datasets built upon the extensive work in the NER community, proposing to recognize the same entities from the same classes but additionally link them to the KB. This perspective is inherited by many EL systems which continue to identify mentions with NER tools. However, the KBs to which EL systems link often contain classes not considered in the MUC-6 consensus.

Hence EL systems began to develop custom ER techniques that target a broader range of entity types. For instance, in Figure 1, "Michael Jackson" would belong to the MUC-6 class *Person*, whereas "Michael Jackson: One" – though present as an entity in KBs such as DBpedia and Wikipedia – would be excluded by MUC-6, referring to a theatrical production. Along these lines, some authors chose to extend the initial MUC-6 classes, including also *Products*, *Financial Entities* [21], *Films*, *Scientists* [6], etc. On the other hand, other authors propose to separate current classes to more specific ones, for instance, deriving *City*, *State*, *Country* from the class *Location* [8]. Generally speaking, however, class-based definitions of entities are inflexible as they cannot hope to adapt to the variety of types present in large KBs. For instance, Wikidata alone has entities from 50,000 unique classes. Therefore, other authors advocate for more general definitions, but these often lack formality [5, 35]. One option is to use a *Miscellaneous* class of entities, but this leaves the question of what sorts of entities this class should cover. Other authors have tried to provide a more general definition of entity, such as the definition "*substrings corresponding to world entities*" used by Ling et al. [17]; however, such a definition is cyclical, due to using the word "*entity*" in the definition of an "*entity*".

Proposal: Instead of addressing the abstract question "*what is an 'entity'?*", our position is to rather address the more practical question "*what should Entity Linking link?*". Posed this way, the question suggests a practical response: "*it depends on the application!*" [32].

Table 1: Survey of popular EL datasets; 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.

Dataset	Languages
AIDA/CoNLL-Complete [11]	EN
KORE50 [13]	EN
IITB [16]	EN
ACE2004 [27]	EN
AQUAINT [27]	EN
MSNBC [3]	EN
DBpedia Spotlight [19]	EN
N3-RSS 500 [28]	EN
Reuters 128 [28]	EN
Wes2015 [36]	EN
News-100 [28]	DE
Thibaudet [1]	FR
Bergson [1]	FR
SemEval 2015 Task 13 [22]	EN,ES,IT
DBpedia Abstracts [2]	DE,EN,ES,FR,IT,JA,NL
MEANTIME [20]	EN,ES,IT,NL
VoxEL$_R$	DE,EN,ES,FR,IT
VoxEL$_S$	DE,EN,ES,FR,IT

For example, in a semantic search scenario, where the goal is to find documents mentioning particular entities or particular types of entities, finding all repeated mentions of an entity may not be so key a requirement for EL: finding any mention in the document might suffice. On the other hand, for relation extraction, each mention might refer to a potential relation in the text, and hence the requirements for EL change. With this in mind, our goals are to first understand on which types of entities there is consensus in the community, and on which not. Then we wish to better understand what are the applications for EL, and how the choice of application affects the requirements of the EL system. Finally we aim to develop EL benchmarks and metrics that – rather than assuming a one-size-fits-all definition of an entity – reflect our findings in terms of varying consensus, applications and requirements.

3.2 Multilingual EL

Thus far, the bulk of effort in EL research has been devoted to English texts. However, more recently, a number of multilingual EL systems – supporting multiple languages – have been proposed. Such systems raise new questions for evaluating EL: How well would state-of-the-art approaches perform over non-English corpora? How would performance vary across languages (and why)? Given that EL often targets named entities, how important is it for EL systems to be configurable for different languages (e.g., *Michael Jackson*'s name does not change with language, only with alphabet)? Given recent improvements in machine translation, how do multilingual EL systems perform versus translating input text?

One challenge faced for responding to these questions is the short list of multilingual datasets available that could be used to

Table 2: Overall EL evaluation (F_1) of selected approaches for the SemEval 2015 Task 13 in Spanish (ES) and English (EN). Approaches configured for Spanish are italicized.

System	ES	EN
Babelfy	0.439	0.602
DBpedia-Spotlight	0.337	0.414
WikiMe	0.043	0.043
TAGME	0.133	0.395
THD	0.069	0.110
AIDA	0.010	0.046

evaluate EL performance for various languages. In Table 1 we survey a variety of EL datasets available in the literature, where we can observe in Table 1 that the majority only consider English text; others that consider non-English texts only offer one language. While a number of multilingual datasets have now been made available, they further present some issues for comparing EL systems across languages, as we will discuss later.

Initial Results: Taking an existing multilingual dataset – SemEval 2015 Task 13 – in [31], we perform initial experiments to compare the performance of popular EL systems for English and Spanish texts, testing Babelfy, DBpedia-Spotlight, WikiMe, TagME, THD and AIDA. Only the first three of these systems can be explicitly configured for Spanish texts. Table 2 offers an overview of the main results. As can be observed, systems generally perform considerably better for English than Spanish, particularly (but not limited to) EL systems not configurable for Spanish. We consider this result as being potentially due to three main factors (a) KBs (e.g., Wikipedia) contains different information for both languages with potentially more information available in English, (b) the models/techniques change according to the target language where, for example, DBpedia-Spotlight's ER uses different models according to the targeted language, and (c) the presence of variations in the languages themselves, where, for example, recognizing "*Star Wars*" is less challenging than the Spanish version "*La guerra de las galaxias*" due to capitalization rules in Spanish and the phrase length.

Further Results: One of the obstacles to ongoing research on multilingual EL is the low availability of datasets with the same text in different languages. According to our survey in Table 2, there are only three multilingual datasets available (the VoxEL datasets are proposed by us). As we detected in the initial work described previously, each of the three datasets has its own limitations. SemEval 2015 Task 13 is composed of four documents on biomedical, math, computer and social topics; DBpedia Abstracts[2] is a large corpus build automatically from the abstracts (first paragraph) the Wikipedia pages, containing in total 39132 documents; and MEANTIME contains annotations of 120 news articles from WikiNews[3] with annotations of entities, events, temporal information and semantic roles. However, DBpedia Abstracts is not a parallel corpus: the text differs across languages, making it unsuitable for comparing performance across languages. On the other hand, while SemEval 2015 Task 13 and MEANTIME aim to be parallel corpora, they have different annotations in different languages. Hence these datasets are not ideal for comparing the performance of EL systems across different languages (though they can be used for comparing EL systems across individual languages). Furthermore, these datasets adopt a particular notion of entity, which as argued previously, may not be that agreed upon by the community.

In order to better compare EL performance across multiple languages, we proposed a new multilingual corpus called VoxEL [30] with the aim of ensuring the same annotations across different languages, as well as reflecting in the dataset the lack of consensus on what is an entity. VoxEL is based on 15 news articles from the European newsletter VoxEurop[4], which is translated by professionals to different European languages. We first aligned the sentences and entities across languages, resolving cases where some entities and sentences were omitted/changed in the translations. To address the lack of consensus about what is an entity, we include two versions of the same dataset: one *strict* version that includes only those entities that all systems appear to agree should be linked (i.e., *Person*, *Location* and *Organization*), and another *relaxed* version that includes also all mentions (including overlapping mentions) with a Wikipedia page related to them (e.g., 'software', 'resort', etc.). Our results again show that – aside from Babelfy – EL systems generally perform much better over English texts. We also compare in [29] the idea of using machine translation in EL environments to translate the input text rather than configuring the EL system for the native text. The results show that the majority of systems – namely DBpedia Spotlight, FREME and TagMe – perform better when the input text is either in English, or translated to English.

3.3 Configurable EL Framework

Our general hypothesis in this Ph.D. work is that when it comes to EL systems, one size does not fit all: different scenarios and different applications may have different requirements for an EL system, including, but not limited to, the types of entities targeted, the languages supported, etc. Drawing the Ph.D. work together, our goal is to be able to perform finer-grained evaluation of EL systems under different requirements and different assumptions. Going one step further, rather than evaluating EL systems, we wish to be able to evaluate the effects of different ER and ED techniques. A given EL system already configures a number of techniques into one solution, where the results of a system then confound the performance of these techniques. While some EL papers present results for ER and ED tasks separately, we propose to go one step further.

Proposal: We propose to create a modular EL framework that implements a selection of the most important EL techniques found in the literature. Such a framework should allow for creating custom EL pipelines that allow for evaluating and composing techniques according to the requirements of a particular application. The challenges posed by this proposal are various and include not only the engineering challenge of developing such a modular framework but also some non-trivial research questions. First we must address the question of how the requirements of a particular application can be represented. Second we must consider what datasets and metrics

[2] http://wiki-link.nlp2rdf.org/abstracts/; January 1st, 2018
[3] https://en.wikinews.org/; January 1st, 2018
[4] http://www.voxeurop.eu/; January 1st, 2018

can be used to evaluate individual EL techniques (and their composition) per these requirements. Though ambitious, if successful, this stage of the Ph.D. work could lead to a better understanding of how EL techniques perform under different assumptions and further offer the practical contribution of a generalized EL framework adaptable to a broader variety of applications and settings.

4 METHODOLOGY

Our next steps are aimed at understanding the consensus in the community regarding what is an "entity" in the context of EL, and defining evaluation protocols (datasets and metrics) accordingly. Further into the future, we aim to start work on the configurable EL framework. More specifically, the steps currently in progress are: (1) Review the state-of-the-art in measures used for EL evaluation. (2) Issue a questionnaire to the EL community to understand the current consensus on the goal of EL systems. (3) Propose a categorization of entities that allows for the inclusion/exclusion of entities per the application requirements. (4) Propose finer-grained evaluation protocols for EL according to previous findings. Later we further wish to being to: (5) Implement a general EL framework based on the previous results that allows us to evaluate and combine techniques according to the application scenario, comparing results against state-of-the-art EL systems in different settings.

5 CONCLUSION

This Ph.D. proposal is motivated by the hypothesis that for EL systems, one size does not fit all. Our first aim has been to understand how the goal of EL systems may vary across different applications and how that affects the consensus of what is an "entity"; this work remains ongoing. Our second aim is to consider how different EL systems perform for different languages, where we have published some results and proposed a novel dataset along these lines. Our third aim is to develop a general EL framework that allows greater configurability for a given application and setting. Our overall ambition is to reach a greater understanding of how EL techniques perform in different settings and arrive at a practical framework that can be used in a broader variety of applications that can, potentially, benefit from the results of the EL task.

Acknowledgments: I wish to thank Aidan Hogan and Barbara Poblete for supervising my PhD research. This work was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017 and by the Millennium Institute for Foundational Research on Data (IMFD).

REFERENCES

[1] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: named entity linking in digital literary editions using linked data sets. *CSIMQ* 7 (2016), 60–80.
[2] Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus.. In *LREC*.
[3] Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP-CoNLL* (2007), 708.
[4] Milan Dojchinovski and Tomás Kliegr. 2013. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In *ECML/PKDD*. 654–658.
[5] Alan Eckhardt, Juraj Hresko, Jan Procházka, and Otakar Smrs. 2014. Entity linking based on the co-occurrence graph and entity probability. In *ERD*. 37–44.
[6] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artif. Intell.* 165, 1 (2005), 91–134.
[7] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*. 1625–1628.
[8] Michael Fleischman. 2001. Automated Subcategorization of Named Entities. In *ACL*. 25–30.
[9] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.
[10] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *COLING*. 466–471.
[11] Johannes Hoffart and et al. 2011. Robust disambiguation of named entities in text. In *EMNLP*. ACL, 782–792.
[12] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[13] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[14] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *EMNLP*. 782–792.
[15] Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. All that Glitters Is Not Gold - Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In *ESWC*. 305–320.
[16] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *SIGKDD*. 457–466.
[17] Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design Challenges for Entity Linking. *TACL* 3 (2015), 315–328.
[18] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. 1–8.
[19] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. ACM, 1–8.
[20] A.L. Minard and et al. 2016. MEANTIME, the NewsReader multilingual event and time corpus. (2016).
[21] Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *LREC*.
[22] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking.. In *SemEval*. 288–297.
[23] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL* 2 (2014), 231–244.
[24] Diego Moussallem, Ricardo Usbeck, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2018. Entity Linking in 40 Languages Using MAG. In *ESWC*. 176–181.
[25] Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2016. J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features. *TACL* 4 (2016), 215–229.
[26] Julien Plu, Giuseppe Rizzo, and Raphaël Troncy. 2016. Enhancing Entity Linking by Combining NER Models. In *ESWC*. 17–32.
[27] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *NAACL-HLT*. 1375–1384.
[28] Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *LREC*. 3529–3533.
[29] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. Machine Translation vs. Multilingual Approaches for Entity Linking. In *ISWC (P&D/Industry/BlueSky)*.
[30] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In *ISWC*. 170–186.
[31] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2017. Multilingual Entity Linking: Comparing English and Spanish. In *LD4IE*. 62–73.
[32] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What Should Entity Linking link?. In *AMW*.
[33] Felix Sasaki, Milan Dojchinovski, and Jan Nehring. 2016. Chainable and Extendable Knowledge Integration Web Services. In *NLP&DBpedia*. 89–101.
[34] Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual Wikification Using Multilingual Embeddings. In *HLT-NAACL*. 589–598.
[35] Victoria S. Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics* 4, 1 (2006), 14–28.
[36] Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Linked data enabled generalized vector space model to improve document retrieval. In *NLP&DBpedia*.
[37] Longyue Wang, Shuo Li, Derek F. Wong, and Lidia S. Chao. 2012. A Joint Chinese Named Entity Recognition and Disambiguation System. In *CIPS-SIGHAN*. 146–151.