Entity Linking with Distributional Semantics

Pablo Gamallo[1] and Marcos Garcia[2⋆]

[1] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es
[2] Grupo LyS, Dep. de Galego-Português, Francês e Linguística
Universidade da Coruña, Galiza, Spain
marcos.garcia.gonzalez@udc.gal

Abstract. Entity Linking (EL) consists in linking name mentions in a given text with their referring entities in external knowledge bases such as DBpedia/Wikipedia. In this paper, we propose an EL approach whose main contribution is to make use of a knowledge base built by means of distributional similarity. More precisely, Wikipedia is transformed into a manageable database structured with similarity relations between entities. Our EL method is focused on a specific task, namely semantic annotation of documents by extracting those relevant terms that are linked to nodes in DBpedia/Wikipedia. The method is currently working for four languages. The Portuguese and English versions have been evaluated and compared against other EL systems, showing competitive range, close to the best systems.

Keywords: Entity Linking, Semantic Annotation, Term Extraction

1 Introduction

Entity Linking (EL) puts in relation mentions of entities within a text with their corresponding entities or concepts in an external knowledge resource. Typically, entity mentions are proper names and domain specific terms which can be linked to Wikipedia pages. Most EL methods include three basic subtasks: i) extraction of the terms likely to be entity mentions in the input text, by using Natural Language Processing (NLP) techniques such as tokenization and multiword extraction; ii) selection of the entity candidates: each mention is associated to a set of entities in the external resource; and iii) selection of the best entity candidate for each mention by making use of disambiguation strategies.

In most cases, two types of approaches are suggested for the selection or disambiguation subtask:

⋆ This research has been partially funded by the Spanish Ministry of Economy and Competitiveness through project FFI2014-51978-C2-1-R

1. *Non-collective approaches*, which resolve one entity mention at each time on the basis of local and contextual features. These approaches generally rely on supervised machine learning models [23, 24, 14, 11].
2. *Collective approaches*, which semantically associate a set of relevant mentions by making use of the conceptual density between entities through graph-based approaches [15, 21, 25, 7, 6, 26, 2, 12, 14, 20, 17, 1, 18].

Many applications can benefit from the EL systems, namely educational applications. Text annotated with EL allows students to have fast access to additional encyclopedic knowledge relevant to the study material, by linking proper names and terms to the corresponding pages in Wikipedia or other external encyclopedic sources. In the research community oriented to educational applications, EL is better known as the task of *semantic annotation* [28]. Given a source text, the semantic annotation task is generally restricted to those mentions in the text referring to the same conceptual category. In fact, the main goal of semantic annotation is to semantically categorize a text by identifying the main concepts or subconcepts the text content is about. As a result, only those mentions that are semantically related are annotated in the text with links (e.g., DBpedia URIs) to their corresponding entities/concepts in an external knowledge database. In [22], the authors describe the *DBpedia Spotlight* system, which can be configured to detect topic pertinence. In order to constrain annotations to topically related entities, a higher threshold for the topic pertinence can be set. This way, texts can be annotated by *DBpedia Spotlight* using semantically related entities.

In this article, we will describe an EL system for the task of semantic annotation. For this specific task, the collective approach, which identifies those mentions associated to conceptually related entities, seems to be the most appropriate strategy.

The main drawback of collective approaches is the fact that the conceptual graph generated is too large and very difficult to explore in an efficient and scalable way. The graph can grow dramatically as the set of entities associated to the different mentions in the text is expanded by making use of different types of semantic relations, including the hierarchical ones (hyperonymy).

To minimize this problem, a collective method is proposed and implemented. Our EL method relies on a distributional similarity strategy to select a restricted set of conceptual relations/arcs between entities. In particular, it only selects relations between the most similar entities. Distributional similarity was computed using Wikipedia articles, as in [8]. The conceptual relations between entities that are not similar in distributional terms are removed from the graph. So, the conceptual graph used to search for the entity candidates is dramatically simplified and, then, can be explored in a more efficient way.

The remainder of the article is organized as follows. In the next section (2), we describe the method: It starts by sketching a brief overview of the proposed strategy. Subsection 2.2 describes how we build an entity database computing distributional similarity. Next, Subection 2.3 is focused on the NLP approaches to term extraction. In Subection 2.4, the entity linking strategy is described.

Then, we evaluate and compare our method in Section 3 and, finally, some conclusions are addressed in Section 4.

2 The Method

2.1 Overview

Our EL method consists of three modules:

Distributional Similarity: This module builds the main encyclopedic resource used by the Entity Linking module. Each Wikipedia entity is put in relation with its most similar entities in terms of distributional similarity. This is the main contribution of our work, since, to our knowledge no EL method relies on such a sort of resource. This is described in Section 2.2.

Term Extraction: This module makes use of NLP strategies to extract the most relevant terms from the text. It is described in Section 2.3.

Entity Linking: This is the core of the system. It makes use of Wikipedia-based resources (such as that built by distributional similarity) and of the terms previously extracted from the text. It consists of two tasks. First, it identifies those relevant terms that are linked to Wikipedia entities and, then, it selects, for each term, the best entity candidate by making use of a disambiguation strategy. This module is described in 2.4.

2.2 Distributional Similarity

We use a distributional similarity strategy to select only semantic relations between very similar entities. This strategy allows us to dramatically simplify the number of relations/arcs to be explored in a collective approach.

Let us see an example. In the English DBpedia, the entity *Aníbal Cavaco Silva* (President of Portugal between 2006-2016) is directly related to 17 categories by means of the hyperonymy relationship: for instance, *Living People*, *Prime Ministers of Portugal*, *People from Loulé Municipality*, etc. If we explore these 17 categories going down to obtain their direct child (or hyponyms), the results are 619,406 new entities, which are in fact co-hyponyms of *Cavaco Silva*. Most of these co-hyponyms have a very vague conceptual relation (e.g. being a living person) with the target entity. In order to remove vague conceptual relations, we only select those entities that can be somehow considered as similar to *Cavaco Silva*. Similarity between two entities is computed by taking into account both the *internal links* appearing in the Wikipedia articles of the two entities, and the set of categories directly classifying them. More precisely, two entities are considered to be similar if they share at least one direct category and a significant amount of internal links.

In our experiments, the target entity *Cavaco Silva* is associated with its most similar entities (first column in Table 1), and for each similar entity we also select the most frequent internal links with which they co-occur (second column of the table). The entities in the second column represent the conceptual context with

regard to which two entities are similar. As a result, we obtain a very restricted and very similar set of entities related to *Cavaco Silva*, which includes other Presidents and Primer Ministers of Portugal. Notice that the target entity is also similar to former Finance Ministers (*Ferreira Leite* and *Vítor Gaspar*), since *Cavaco Silva* also had that political function before becoming Prime Minister. In addition, he shares with these two individual the fact of being Economist and having been working at the same universities.

In our experiments, both *similar* and *contextual* entities are all considered in the same way: all are directly related to the target entity. As the list of co-hyponyms for each entity is reduced from some hundred thousands candidates to a few entities (similar and contextual ones), the resulting database is easy to explore by most searching strategies.

Table 1. Entities related to *Aníbal Cavaco Silva* using distributional similarity

similar entities	**contextual entities**
Mário Soares	President of Portugal, Ordem Nacional do Cruzeiro do Sul
	Ordem do Libertador
Jorge Sampaio	António Guterres, Timor-Leste Portuguese Presidential Election
	Ordem de Amílcar Cabral
Diego Freitas do Amaral	Prime Minister of Portugal, New University of Lisbon
	Catholic University of Portugal
Manuela Ferreira Leite	National Assembly of the Republic, Economist, Bank of Portugal
	Fundação Calouste Gulbenkian
Vítor Gaspar	Economist, Francico Louçã, Bank of Portugal, Professor

Let e_1 and e_2 be two entities with the corresponding articles in Wikipedia. They are comparable if they share at least one Wikipedia category. Distributional similarity is only computed on entity pairs sharing at least one category. So, if entities e_1 and e_2 share at least one category, they are actually comparable and similarity is computed. Distributional similarity is computed using the following version of the *Dice* coefficient [3] :

$$Dice(e_1, e_2) = \frac{2 * \sum_i min(f(e_1, link_i), f(e_2, link_i))}{f(e_1) + f(e_2)} \quad (1)$$

where $f(e_1, link_i)$ represents the number of times the entity e_1 co-occurs with the internal link $link_i$. Internal links stand for the distributional contexts of the compared entities. As a result, each entity is assigned a set of similar entities ranked by Dice similarity and a set of internal links ranked by frequency. The resulting entity database is the main knowledge base considered by our semantic annotation strategy. This resource is called *Similarity Knowledge Base*. In [9], *Dice* turned out to be one of the most reliable similarity measures for distributional semantics.

2.3 NLP Techniques for Term Extraction

We distinguish two different types of terms: *basic terms* and *multiword expressions*. Basic terms are lexical units codified as common nouns, adjectives,

verbs, or proper names which are considered as relevant for a given text. Except proper names, which can be composite expressions (e.g., New York, University of South California), basic terms are just single words. Multiwords are relevant expressions codified as compounds that instantiate specific patterns of PoS tags. For instance, `discussion forums`, `natural language`, `cells of plants` or `professor at New University of Lisbon` can be multiwords within a text.

For the specific task of semantic annotation, we assume that not all terms within a text which are linked to an entity (or concept) in DBpedia are semantically relevant. There are frequent mentions, e.g. `concept`, `term`, `red`, etc, which are linked to concepts in DBpedia, but which may not be relevant in some texts. So, terms must be ranked according to their relevance in a text and should be considered as entity candidates only the most relevant ones.

Our approach to extract basic terms and multiwords requires PoS tagging, which is performed with the multilingual NLP suite *CitiusTool* [10].[3] For extracting basic terms, we use a different strategy that the one used for multiword extraction. The strategy we follow to extract basic terms is slightly different from that used for multiwords. In the case of basic terms, their extraction relies on the notion of *termhood*, that is, the the degree that a linguistic unit is related to domain-specific concepts [19]. In the case of multiwords, the extraction is based on the notion of *unithood*, which concerns with whether or not sequences of words should be combined to form more stable lexical units. More formally, unithood refers to "the degree of strength or stability of syntagmatic combinations and collocations" [19]. The concept of unithood is only relevant to complex units (multiwords).

Extraction of Basic Terms The first step consists in identifying and selecting common nouns, adjectives, verbs, and proper names from a given text. Proper names are selected by using named entity recognition. The result is a list of term candidates.

The second step consists in providing the term candidates with a statistical weight, representing the conceptual relevance of the term within the input text. The weight of a term is computed by considering the frequency observed in the input text (observed data) with regard to its frequency in a large collection of texts taken as a corpus of reference (expected data). More precisely, the weight of a term is the *chi-square* value, which measures the divergence between the observed data and the values that would be expected. Expected values are provided by the reference corpus. Finally, all weighted terms are ranked according to their score and the N most relevant are selected for semantic annotation. This way, terms very frequent in the reference corpus (common concepts such as for instance `person`, `thing`, `object`, etc.) tend to be assigned low chi-square values. By contrast, very frequent terms in the input text but rare in the reference corpus have high values and, then, are considered as relevant for the given text.

[3] Freely available at `http://gramatica.usc.es/pln/tools/CitiusTools.html`

Multiword Extraction The proposed strategy relies on the notion of unithood and has common aspects with similar work requiring linguistic patterns [29, 27]. Our extraction of multiwords also consits of two steps: candidates selection and statistical ranking. In the first step, candidates are extracted using a set of patterns of PoS tags. This is the set we use for our four languages:

$noun - adj$	$adj - noun$
$noun - noun$	$noun - prep - noun$
$noun - prep - adj - noun$	$noun - prep - noun - adj$
$adj - noun - prep - noun$	$noun - adj - prep - noun$
$adj - noun - prep - noun - adj$	$noun - adj - prep - noun - adj$
$adj - noun - prep - adj - noun$	$noun - adj - prep - adj - noun$

In the second step, the candidates are ranked according to the notion of unithood: A lexical measure, *chi-square*, provides a test of association between the constituents of a multiword, in order to verify whether the constituents are or are not put together by random . More precisely, the observed values of a multiword stands for its frequency in the input text, while the expected values are derived from the single occurrences of its constituents in the same text.

2.4 The Entity Linking Strategy

Resources and Terms Our strategy makes use of three resources, which represent three different linguistic relations:

Similarity Knowledge Base (SIM) This stands for similarity relationships between Wikipedia entities. Wikipedia entities correspond to the titles of articles in Wikipedia (dump file of December 2014). This resource was built based on distributional similarity (see Section 2.2 above).[4]

Categories of Wikipedia entities (HYPER) This database contains hierarchical (hyperonymy) relations between Wikipedia entities and their direct parent categories. This resource is provided by DBpedia[5].

Redirects of Wikipedia entities (REDIR) This database contains synonymous relations between Wikipedia entities and their different names. This resource is also provided by DBpedia.

The union of Wikipedia entities and categories gives rise to the set of (conceptual) entities of our ontology. Indeed, some categories are not Wikipedia entities.

Besides these three resources, our EL strategy also relies on term extraction (see Section 2.3). The output of this task, which is a ranked list of relevant terms (both single words and multiwords) is the input of the following EL tasks: searching for candidates and disambiguation. According to [13], the most efficient EL systems divide the process of entity linking in these two tasks. During the search phase the system proposes a set of candidates for an entity mention to be linked to, which are then ranked by the disambiguator.

[4] This resource is available from the authors upon request.

[5] `http://downloads.dbpedia.org/3.8/`

Searching for Entity Candidates We verify whether the relevant terms extracted from the input text are actually mentions of entities. For this purpose, they are expanded in two different ways: 1) Each term is expanded with its lemma, for instance the term `databases` is expanded with the singular form `database`. 2) Terms are expanded with their synonymous stored in the resource REDIR. All the inflected forms and synonyms of a term occurring in the input text are joined in a single terminological unit. Then, we search for semantic links between expanded terms (terminological units) and entities. The search for links between terms and entities is performed using our external resources: SIM, REDIR, and HYPER. The main problem arising when terms are intended to be linked to entities is term ambiguity.

One term (hereafter we use interchangeably "term" and "terminological unit") can be associated to several entities, which represent their different senses. A natural way of accessing the different entities/senses of an ambiguous term is to use Wikipedia disambiguation pages. However, these pages include many odd senses which should not be linked to the ambiguous term. For instance, the French town *Barcelonnette* is considered as one of the senses of the term `Barcelona`, which is clearly odd. Instead of using the entities listed in the disambiguation pages, we select the entities/senses of an ambiguous term by taking into account some regular expressions related to the syntax of the Wikipedia titles. In Wikipedia, different entities with the same name are individualized by making use of brackets, commas or hyphens. For instance, the ambiguous term `Paris` is associated to entities like *Paris*, *Paris,_Ohio*, *Paris,_Arkansas*, *Paris_(mythology)*, *Paris_(song)*, etc. All of them can be considered different senses of the original term. Even if our use of regular expressions in Wikipedia titles does not always include all possible senses of an ambiguous term, most extracted senses are apparently correct ones. So, our technique is more precise than that based on disambiguation pages but has lower coverage.

The output of this task is a list of entity candidates associated with all relevant terms extracted from the input text.

Weighting Candidates and Entity Disambiguation In this task, we select the best entity candidate of each term by making use of a disambiguation strategy. This strategy relies on selecting the entity with the highest weight for each term.

Given a term, the process starts by assigning the same weight to all its entity candidates. Then, it explores the semantic relationships (similarity and hyperonymy) of each entity candidate and searches for related entities that are also semantically related to the candidates of the other terms in the input text. The procedure of exploring and searching common related entities is performed on the two knowledge resources: SIM (similarity) and HYPER (hyperonymy).

The weighting process is just a summatory of semantically related entities that are shared by both the target entity and the rest of entity candidates of all input terms. Given a terminological unit t_1 and an entity candidate e_1, the final weight of this entity with regard to t_1 is computed as follows:

$$weight(t_1, e_1) = \sum_{i=1}^{k} sim(e_1, e_i) + hyper(e_1, e_i) \quad (2)$$

where $sim(e_1, e_i)$ stands for the number of similar entities which are shared by e_1 and each member (e_i) of the set of entity candidates; $hyper(e_1, e_j)$ represents the number of categories which are shared by e_1 and each member of the set of entity candidates. The former function is computed on SIM while the latter works on HYPER. The set of entity candidates is constituted by those entities associated to all terminological units extracted from the text, where k is the size of the set. Finally, for each term, the entity with the highest *weight* value is selected.

Let us take an example. Suppose we have selected the term `Cavaco Silva`. To compute *weight*(`Cavaco Silva`, *Aníbal Cavaco Silva*) given the pool of entities {*Aníbal Cavaco Silva*, *Jorge Sampaio*, *Lisbon*}, we compute first the *sim* function, which consists in counting the number of entities in the pool which are similar to the target entity *Aníbal Cavaco Silva* according to the SIM database. Given the table 1 above, only one of these entities is linked by similarity to the target entity. So, the result of the *sim* function is just 1. A similar procedure is performed to comput the *hyper* value, but using the HYPER resource.

System Implementation The method was implemented in Perl giving rise to the system called *CitiusLinker*. So far, it works for four languages: English, Portuguese, Spanish, and Galician.[6] In order to facilitate its integration into external web processes, we also implemented a RESTful web service with Dancer.[7] The web service interface can be used to annotate the text with the selected terms and their linked entities. Besides, it also gives as output a set of semantically related DBpedia entities to those found in the text (semantic enrichment), as well as a set of DBpedia categories that can be used to classify the text (semantic categorization). The web service returns HTML, XML, YAML or JSON output documents. It can be configured to select one of the four languages, the output format, and the number of relevant basic terms.

3 Evaluation

In order to provide an evaluation of our system in the task of semantic annotation, we performed two experiments with English and Portuguese texts, using manually annotated test corpora.

For English, we used the DBpedia Spotlight Evaluation Dataset [22]. The test corpus consists of 10 randomly selected excerpts from New York Times news, and each excerpt/document was manually annotated with DBpedia concepts. For

[6] A demo is available at `http://fegalaz.usc.es/~gamallo/demos/semantic-demo/`
[7] `http://fegalaz.usc.es/nlpapi`

Portuguese, we created a similar dataset from 10 different *Jornal de Notícias* news, which were manually annotated by two linguists using the Portuguese DBpedia. To build the gold standard dataset, we selected the concepts identified by both annotators. As a result, we obtained 130 concepts for the 10 documents.

Both annotated datasets are freely available.[8]

Notice that the evaluated task is different from that defined in the different TAC-KBP Entity Linking Tracks [16]. In those tracks, the objective is not to identify the relevant concepts of a given document, but identifying the correct node/concept in DBpedia given a name mention in a document. Besides, the test datasets are just focused on named entities of type PER (person), ORG (organization), or GPE (geopolitical entity). In [5], the author describes the construction of two datasets for entity linking in the Portuguese and Spanish languages, by making use of the cross-lingual XLEL-21 dataset. This dataset is equivalent to the one used in TAC-KBP, and contains just person names.

In the English evaluation, we compare our results with those of several publicly available annotation services. The results of all systems were obtained by using the same gold standard: DBpedia Spotlight Evaluation Dataset. Except *CitiusLinker* and *Alchemy*, whose F_1 scores were obtained from our own experiments, the scores of the remainder systems were taken from [22].

Table 2. F_1 scores reached by different EL systems using the DBpedia Spotlight Evaluation Dataset (for English)

Systems	**F_1-score**
The Wiki Machine[a]	59.5%
DBpedia Spotlight (best configuration)	56.0%
CitiusLinker (best configuration)	55.9%
Zemanta[b]	39.1%
Alchemy[c]	21.1%
Open Calais[d]	14.7%
Ontos[e]	10.6%

[a] `http://thewikimachine.fbk.eu`
[b] `http://www.zemanta.com`
[c] `http://www.alchemyapi.com`
[d] `http://www.opencalais.com`
[e] `http://www.ontos.com`

Table 2 shows that the performance of our strategy, *CitiusLinker*, is in a competitive range for English, close to the two best systems: *Wiki Machine* and *DBpedia Spotlight*.

Concerning the Portuguese evaluation, results are depicted in Table 3. Unfortunately, we only could compare our system to DBpedia Spotlight and that

[8] `http://gramatica.usc.es/~gamallo/datasets/el_dataset.tar.gz`

Table 3. F_1 scores reached by three systems using the Portuguese dataset

Systems	**Precision**	**Recall**	**F_1-score**
CitiusLinker (best configuration)	45.3%	56.2%	50.9%
DBpedia Spotlight (best configuration)	45.6%	51.2%	48.4%
Alchemy	12.8%	5.38%	7.56%

provided by *Alchemy*. To the best of our knowledge, no further EL systems for Portuguese are available yet. The scores reached by *CitiusLinker* and *DBpedia Spotlight* are slightly lower than those got in the English evaluation. Both systems achieve similar F_1-score values after having set their parameters to find the best configuration. By contrast, *Alchemy* system dramatically drops performance. In this case, no parameter configuration has been done since the experiments were performed from the API server provided by the company. The Portuguese *DBpedia Spotlight* version belongs to a multilingual system which is described in [4].

The F_1-score of our system has been obtained with the best configuration: 60 most relevant basic terms (only nouns) and all multiwords. When using adjectives and verbs, the F_1-score decreases. Notice also that no multiword was filtered out. Unlike basic terms, which can refer to very generic concepts in some cases, multiwords linked to DBpedia entities are likely to be domain-specific terminological expressions referring to specific concepts. By default, *CitiusLinker* selects all multiwords found in the text.

4 Conclusions

In this article, we proposed a method for a specific entity linking subtask, namely semantic annotation with DBpedia concepts. The main contribution of our method is the use of an external entity base built by means of distributional similarity. This entity base is structured with similarity relationships between entities which are not directly related by means of the DBpedia resources. In the disambiguation process, our method only explores the similarity relations found in this entity base, as well as the direct hyperonymy relationships provided by DBpedia. This way, the weighting process used to disambiguate becomes simpler and more efficient than those based on exploring several levels of organization through DBpedia or any other ontology. Another important contribution of our method is the use of different NLP techniques for term extraction. We defined a specific strategy for the extraction of basic terms, which is different from multiword extraction. Our approach achieved competitive performance over the traditional methods in English, while kept similar performance in Portuguese. In future work, we will evaluate the results obtained for languages other than English and Portuguese. A deep qualitative error analysis is also required in order to find the main drawbacks of our approach. It will also be adapted to be applied on TAC-KBP tasks in order to be compared to other EL systems.

References

1. Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, and Hongzhao Huang. Analysis and Enhancement of Wikification for Microblogs with Context Expansion. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Technical Papers*, pages 441–456, 2012.
2. Silviu Cucerzan. TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation. In *Proceedings of the Text Analysis Conference (TAC 2011)*, 2011.
3. James R. Curran and Marc Moens. Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, volume 9, pages 59–66, Philadelphia, 2002.
4. Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, pages 121–124. Association for Computing Machinery, 2013.
5. João Tiago Luís dos Santos. *Linking entities to Wikipedia documents*. PhD thesis, Instituto Superior Técnico, Lisboa, 2013.
6. Norberto Fernández, Jesus A. Fisteus, Luis Sánchez, and Eduardo Martín. WebTLab: A cooccurrence-based approach to KBP 2010 Entity-Linking task. In *Proceedings of the Text Analysis Conference (TAC 2010)*, 2010.
7. Paolo Ferragina and Ugo Scaiella. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 1625–1628, Toronto, 2010.
8. Pablo Gamallo. Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. In *LREC 2008 Workshop on Comparable Corpora*, pages 19–26, Marrakesh, 2008.
9. Pablo Gamallo and Isaac González. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71, 2011.
10. Marcos Garcia and Pablo Gamallo. Yet another suite of multilingual nlp tools. In *Symposium on Languages, Applications and Technologies (SLATE 2015)*, pages 81–90, Madrid, 2015.
11. Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1020–1030, 2013.
12. Yuhang Guo, Wanxiang Che, Ting Liu, and Sheng Li. A graph-based method for entity linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 1010–1018, 2011.
13. Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence*, 194:130–150, 2013.
14. Xianpei Han and Le Sun. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11)*, volume 1, pages 945–954, Portland, Oregon, 2011.
15. Xianpei Han and Jun Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pages 215–224, Hong Kong, China, 2009.

16. Joel Nothman Heng Ji and Ben Hachey. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proceedings of the Text Analysis Conference (TAC 2014)*, pages 539–545, 2014.
17. Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 782–792, 2011.
18. Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. Collective Tweet Wikification based on Semi-supervised Graph Regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Volume 1: Long Papers*, pages 380–390. Association for Computational Linguistics, 2014.
19. Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology*, 3(1):259–289, 1996.
20. Zornitsa Kozareva, Konstantin Voevodski, and Shang-Hua Teng. Class label enhancement via related instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 118–128. Association for Computational Linguistics, 2011.
21. Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 457–466, Paris, 2009. Association for Computing Machinery.
22. Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, Graz, 2011. Association for Computing Machinery.
23. R. Mihalcea and A. Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, pages 233–242, Lisbon, 2007.
24. D. Milne and I.H. Witten. Learning to Link with Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'08)*, pages 509–518, Napa Valley, 2008.
25. Marco Pennacchiotti and Patrick Pantel. Entity extraction via ensemble semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 238–247, 2009.
26. Will Radford, Ben Hachey, Joel Nothman, Matthew Honnibal, and James R. Curran. Document-level Entity Linking: CMCRC at TAC 2010. In *Proceedings of the Text Analysis Conference (TAC 2010)*, 2010.
27. D. Sánchez and A. Moren. A methodology for knowledge acquisition from the web. *Journal of Knowledge-Based and Intelligent Engineering Systems*, 10(6):453–475, 2006.
28. Juan C. Vidal, Manuel Lama, Estefanía Otero-García, and Alberto Bugarín. Graph-based semantic annotation for enriching educational content with linked data. *Knowledge-Based Systems*, 55:29–42, 2014.
29. Jordi Vivaldi and Horacio Rodríguez. Improving term extraction by combining different techniques. *Terminology*, 7(1):31–47, 2001.