JeuxDeLiens: Word Embeddings and Path-Based Similarity for Entity Linking using the French JeuxDeMots Lexical Semantic Network

Julien Plu[1] Kévin Cousot[2]
Mathieu Lafourcade[2] Raphaël Troncy[1]
Giuseppe Rizzo[3]
(1) EURECOM, Sophia-Antipolis, France
(2) LIRMM, Montpellier, France
(3) ISMB, Turin, Italy
julien.plu@eurecom.fr, kevin.cousot@lirmm.fr
mathieu.lafourcade@lirmm.fr, raphael.troncy@eurecom.fr
giuseppe.rizzo@ismb.it

RÉSUMÉ

Les systèmes de désambiguïsation d'entités nommées utilisent principalement sur des bases de connaissances encyclopédique telles que DBpedia ou Freebase. Dans ce papier, nous utilisons à la place, un réseau léxicalo-sémantique nommé JeuxDeMots pour conjointement désambiguïser et typer les entités nommées. Notre approche combine les plongements de mots et la similitude de chemins dans un graphe résultant à des résultats encourageants sur un ensemble de documents provenant du journal Le Monde.

ABSTRACT

Entity linking systems typically rely on encyclopedic knowledge bases such as DBpedia or Freebase. In this paper, we use, instead, a French lexical-semantic network named JeuxDeMots to jointly type and link entities. Our approach combines word embeddings and a path-based similarity resulting in encouraging results over a set of documents from the French Le Monde newspaper.

MOTS-CLÉS : entité nommée, désambiguïsation, jeuxdemots, réseau lexical.

KEYWORDS: named entity, disambiguation, jeuxdemots, lexical network.

1 Introduction

There is an exponential growth of textual content made available on the Web, produced by anyone via a broad diversity of publishing platforms. Automated solutions to extract actionable insights from this content is therefore of utmost importance. Focusing on textual content, we identified four main challenges that the NLP community is tackling for performing an entity linking task : dealing with different genres (social media, video subtitles, newswire articles, search queries), written in different languages, mentioning entities related to a variety of domains that can be classified with diverse sets of per-domain classes, and disambiguated against multiple kind of datasources (knowledge bases,

relational databases, lexical-semantic networks) (Plu, 2016).

Information extraction aims to get structured information from unstructured text by attempting to interpret natural language text to extracting information about entities, relations among entities and linking entities to external referents. In detail, entity recognition aims to locate and classify entities in text into defined classes such as Person, Location or Organization. Entity linking (or entity disambiguation) aims to disambiguate entities in text to their corresponding counterpart, referred as resource, contained in a knowledge base. Each resource represents a real world entity with a specific identifier.

Many knowledge bases can be used for doing entity linking : DBpedia (DBpedia, 2007), Freebase (Freebase, 2007), Wikidata (Wikidata, 2012) to name a few. Those knowledge bases are known for being broad in terms of coverage, while vertical knowledge bases also exist in specific domains, such as Geonames (Geonames, 2006), Musicbrainz (MusicBrainz, 2000), LinkedMDB (LinkedMDB, 2009), and here JeuxDeMots.

We have decided to focus our work on linking entities using a lexical-semantic network as a referent base. We have identified four main lexical-semantic networks for the French language : Wolf (Sagot & Fier, 2008), JeuxDeMots (Lafourcade, 2007b), FLN (Lux-Pogodalla & Polguère, 2011) and Babelnet (Navigli & Ponzetto, 2012), the latter being multilingual. We have selected the JeuxDeMots network (Lafourcade, 2007a) as it offers daily updates and it is well-suited for disambiguation thanks to its polysemy representation. This network is an oriented graph where the nodes can be labeled with terms, concepts or any kind of textual item. The edges are oriented, weighted and indicate a specific type of relation between two vertices. They can be lexical (lemma, locution, action to verb, etc.) or semantic (hypernymy, meronymy, agent, etc.). So far, the JeuxDeMots network has more than 1.2 million nodes, 67 million edges and about 100 relationship types.

The reminder of the paper is structured as follows : Section 2 detail the relevant previous work already done. Section 3 introduces our two steps approach for entity linking. Next, the Section 4 proposes an evaluation of this approach. Finally, we conclude and discuss some future work in Section 5.

2 Related Work

Currently, for the French language, the state of the art for entity linking suffers from a lack of evaluated approaches, resources and datasets. One of the most recent entity linking method (Stern *et al.*, 2012) relies on a statistical model trained on a manually annotated corpus and uses features computed using a basic heuristic tool and extracted from a large knowledge base. Unfortunately, this method is hard to reproduce since the knowledge base and the corpus used in the experiments are not publicly available. We describe two methods that can be applied to French : Babelfy (Moro *et al.*, 2014) and DBpedia Spotlight (Daiber *et al.*, 2013).

Babelfy proposes a graph-based approach where two main algorithms have been developed : random walk and a heuristic for finding the subgraph that contains most of the relations between the recognized mentions and candidates. The nodes are pairs (mention,entity) and the edges correspond to existing relationships in BabelNet (Navigli & Ponzetto, 2012) that are scored. Next, a semantic graph is built using word sense disambiguation (WSD) that extracts lexicographic concepts and does entity linking by matching strings with resources described in a knowledge base.

In contrast, DBpedia Spotlight relies on the so-called TF*ICF (Term Frequency-Inverse Candidate Frequency) score computed for each entity. The goal of this score is to show that the discriminative strength of a mention is inversely proportional to the number of candidates it is associated with. This means that a mention that commonly co-occurs with many candidates is less discriminative. Although, those methods can be applied on French documents, they have never been thoroughly evaluated with this language due to a lack of proper benchmark datasets. Next, those methods do not take into account the different semantics that an entity might have (see in Section 3 an example with *Paris*).

3 JeuxDeLiens Approach

We have originally developed a named entity recognition (NER) model trained with the ETAPE (Gravier *et al.*, 2012) corpus for recognizing entities in French documents. However, the corpus required a lot of pre-processing that was consisting of cleaning the transcripts (speech-to-text dataset). In fact, even after the cleaning the performance of the NER was very low. The extraction part not being our focus, we have assumed to take as input the surface forms of the entities that have to be linked and typed with the JeuxDeMots network.

Our entity linking approach is designed in two steps : *i)* word embeddings, and *ii)* path-based similarity. As a running example, we will use the following three sentences :

— Paris compte au 1er janvier 2013 plus de 2,2 millions d'habitants. [1] *Paris* stands for *the French capital*.
— En matière de commerce, Paris a clairement affiché sa volonté. [2] *Paris* stands for *the French government*.
— Paris fait ses premires apparitions comme mannequin dans plusieurs évènements de charité. [3] *Paris* here for *Paris Hilton*.

Entities are annotated beforehand in the text between double square brackets (e.g. [[Paris]]). We perform a preprocessing that consists in the following three steps :

1. tokenizing the input text by respecting the annotated entities between double square brackets ;
2. once the text is properly tokenized, we link each word to a node in the JeuxDeMots network, and we generate all the possible entity candidates. In the three sentences, the candidate entities for Paris are : *i)* Paris Hilton, *ii)* Paris, capitale [4], *iii)* Paris, gouvernement français [5], *iv)* Paris, prénom [6], and *v)* Paris, nom [7]. We also map the words that are multitokens expression such as *table de chevet* [8]. Finally, if no entity candidates are found, we assume the entity to be a novel entity [9] and will be linked to NIL ;
3. once the mapping is done, we remove the stopwords.

After this preprocessing, we have a list of words that represent the context of the document, and where each of them has its node mapped in the JeuxDeMots network. The next step is to select the best entity candidate in the list and we use a word2vec (Mikolov *et al.*, 2013) embedding.

1. As of 1 January 2013, Paris has more than 2.2 million inhabitants
2. In terms of trade, Paris has clearly stated its will.
3. Paris made her first appearances as a model in several charity events.
4. capital city
5. French government
6. firstname
7. familyname
8. nightstand
9. entities that do not appear in the data source being used

3.1 Word Embeddings

In order to properly disambiguate the entities, we use word2vec as word embeddings method with a model trained over the frWac corpus (Baroni *et al.*, 2009). This method has been chosen since a surface form might have more than one meaning as shown in our running example. A simple string comparison between surface forms, e.g. using the Levenshtein distance, is not efficient since the two contexts, of the entity and into JeuxDeMots shall be compared. In JeuxDeMots, the context of an entity is represented by his gloss (e.g. gouvernement français, prénom, capitale, etc.). For the first sentence, instead of comparing the surface form *Paris* with the words of the context, we use *gouvernement français* and the others gloss such as in Equation 1. The set W represents the following context : [compte, janvier, millions, habitants].

$$\begin{aligned} &w2v([gouvernement, francais], W) \\ &w2v([capitale], W) \\ &w2v([Paris, Hilton], W) \\ &w2v([prénom], W) \\ &w2v([nom], W) \end{aligned} \tag{1}$$

The entity candidate that got the best score is taken as the proper linked entity. We also propose to group the novel (*NIL*) entities that may identify the same real-world thing. We attach the same *NIL* value within and across documents. For example, if we take two different documents that share the same emergent entity, this entity will be linked to the same *NIL* value. We can then imagine different *NIL* values, such as *NIL_1*, *NIL_2*. We perform a string matching over the surface form between each novel entities that have been linked to *NIL* (or between each token if it is a multiple token mention). Concerning our running example, with this approach, all the *Paris* entities have been properly linked to their corresponding meaning. We use a path-based similarity method to find the type of the entity.

3.2 Path-Based Similarity

At this stage, we need to assign a type to an entity. From the JeuxDeMots network, we can gather the context words W, the entity e and the set of possible classes C. Although, we got W during the mapping step and e with word2vec, the class nodes are simply hand-picked nodes describing the class. Here, we only use *lieu* [10] for LOC, *organisation* for ORG and *personne* [11] for PER as classes.

The logic is to find the best path in the JeuxDeMots network from W to each element of C passing by e. An example is represented in Figure 1 for the first sentence. In case e is a novel entity, we try to directly find the best path from W to each element of C. The JeuxDeMots network has two characteristics that we have to take into account : *1)* a direct link from the entity to the class might not exist, there could be an indirect link (*Paris (capitale)* $\xrightarrow{\text{is-a}}$ métropole [12] $\xrightarrow{\text{is-a}}$ lieu), or multiple links (*Paris (capitale)* $\xrightarrow{\text{is-a}}$ {ville [13], préfecture [14], destination touristique [15]} and *2)* the JeuxDeMots taxonomy has no constraints, has multiple words to express the same meaning of a class and has loops.

10. place
11. person
12. metropolis
13. city
14. prefecture
15. touristic destination

In order to choose the class of the entities, we propose a graph similarity measure. Similarity measures are used in information retrieval (Franzoni *et al.*, 2014), semantic path analysis (Song *et al.*, 2015) and link prediction (Lü *et al.*, 2009). Some of them only take nodes attributes into account while others are based on nodes neighborhood or even paths (Liben-Nowell & Kleinberg, 2007). We propose a path-based similarity measure inspired from LP-index (Lü *et al.*, 2009). First, the score of a path μ is defined as :

$$score(\mu) = \delta^{|\mu|-1}. \sum_{\forall e=(u,v)\in\mu} w_e.|R^+_{T_e}(u)|^{-1} \tag{2}$$

With $\delta \in [0, 1]$ a length-malus, $|\mu|$ the length of the path, $R^+_{T_e}(u)$ the set of outgoing relations of u with the same type T_e as e and w_e its weight.

The similarity between two nodes u and v can then be expressed as :

$$sim(u, v) = \frac{1}{|M_{u,v}|}. \sum_{\forall \mu\in M_{u,v}} score(\mu) \tag{3}$$

With $M_{u,v}$ the set of paths between u and v. If the entity was found in the network we choose the class $c = \arg\max_{c\in C} sim_{ctx}(W, e, c)$ with :

$$sim_{ctx}(W, e, c) = sim_s(W, e) + sim(u, c) \tag{4}$$

And :

$$sim_s(W, e) = \frac{1}{|W|}. \sum_{\forall w\in W} sim(w, e) \tag{5}$$

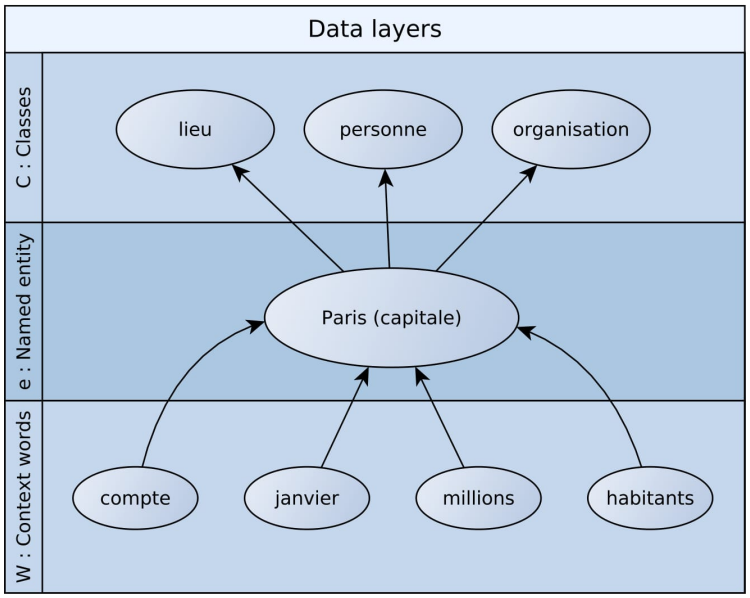

FIGURE 1 – The 3 data layers : classes, named entity and context words.

If the network does not contain the entity, then we only use its context to try and choose the class $c = \arg\max_{c\in C} sim_s(W, c)$. In practice, we limit the length of the paths to 2 and only the relevant types of relations are considered.

4 Evaluation

To evaluate JeuxDeLiens, we have selected 15 articles from the French newspaper Le Monde. We were not able to compare our approach with the ones used in Babelfly and DBpedia Spotlight. In fact, the used algorithms cannot be applied on the JeuxDeMots network (see Section 2) without changing their core.

The French community suffers from a lack of datasets for training entity linking systems. Therefore, we have decided to build a dataset based on Le Monde newspaper articles. We have randomly selected 15 articles that we have manually annotated. Only the entities that represent a *Person*, a *Location* or an *Organization* have been annotated, and linked to their JeuxDeMots ID if they exists in the lexical-semantic graph, otherwise they have been linked to NIL. The NIL entities keep the same ID across the documents, in order to respect the NIL clustering logic detailed in Section 3.1. The tool BRAT [16] has been used for the annotation task. In order to create a dataset with proper statistics, we have decided to align the numbers as much as possible over the ones from OKE Challenge 2015 Task 1 (Nuzzolese *et al.*, 2015) test dataset. We put together the statistics of these two datasets in Table 1 to be able to compare them.

As scorer, we use the neleval scorer (Hachey *et al.*, 2014) with the following metrics :

- *strong_typed_all_match* : performs a micro-averaged evaluation of all mentions. A mention is counted as correct if it is a correct link or a correct nil. A correct link must have the same span, entity type, and KB identifier as a gold link. A correct NIL must have the same span as a gold NIL.
- *entity_ceaf* : performs an evaluation based on a one-to-one alignment between system and gold entity clusters for both KB identifier and NIL across documents.
- *strong_typed_mention_match* : performs a micro-averaged evaluation of entity mentions. A system span must match a gold span exactly to be counted as correct and additionally requires the correct entity type.
- *strong_link_match* : performs a micro-averaged evaluation of links. A system link must have the same span and KB identifier as a gold link to be counted as correct.
- *strong_nil_match* : performs a micro-averaged evaluation of NIL entities. A system NIL must have the same span as a gold NIL to be counted as correct.
- *strong_all_match* : performs a micro-averaged link evaluation of all mentions. A mention is counted as correct if is either a link match or a NIL match as defined above.

As we can see, we evaluate JeuxDeLiens at multiple level in order to identify its strengths and weaknesses. The results are shown in Table 2. We also remind that we start from a score of 100% in F1 at extraction level as we take as input the correct surface forms of the entities.

	OKE2015	JeuxDeLiens
Number of unique entities of type PERSON	155	152
Number of unique entities of type LOCATION	90	102
Number of unique entities of type ORGANIZATION	122	131
Number of entities of type PERSON	317	228
Number of entities of type LOCATION	123	117
Number of entities of type ORGANIZATION	141	226

TABLE 1 – Statistics over the dataset

16. http://brat.nlplab.org/

	Precision	Recall	F1
strong_typed_all_match	63.2	63.2	63.2
entity_ceaf	73.3	93.3	82.1
strong_typed_mention_match	64.9	64.9	64.9
strong_link_match	72.9	89.7	80.5
strong_nil_match	100	50	66.7
strong_all_match	77.2	77.2	77.2

TABLE 2 – Results for JeuxDeLiens

From the error analysis conducted, it emerges that the heterogeneity of the network plays an important role for entity linking. Indeed, the knowledge is unevenly distributed across the network and while some domains are well-supplied, some are not. The entity coverage is very satisfying (89.7% recall for the *strong_link_match* score) but some entities have a very few incident edges making a path difficult to find. This is the case for *Boston Dynamics* and *BigDog*. Symmetrically, human-related nodes (*man*, *woman*, *human*...) are heavily well-supplied. As they are high-degree nodes, more paths are found leading to more misclassification favoring *PER*. This explains why some entities have been mistyped. The *entity_ceaf* score shows that the approach can succeed to give the same link for the entities that have the same meaning, including the ones that have been linked to same *NIL* across the documents. Nevertheless, the precision is a bit low because of a weakness of our *NIL* clustering method that might link unrelated entities if they share a same token in their mention. The score *strong_nil_match* reveals that when we link an entity to *NIL*, it is a good guess. However, our system still proposes some candidates for some entities that should be linked to *NIL*. Our word2vec measure has also a problem when processing a word that does not belong to the vocabulary since no similarity is computed. This impacts emergent entities that are present in JeuxDeMots but not in the frWac corpus.

5 Conclusion and Future Work

We have proposed JeuxDeLiens, a two steps method for disambiguating entities in French documents using the JeuxDeMots lexical semantic network. Although the evaluation has been made on a single newspaper article, the results of JeuxDeLiens are very encouraging and show that our approach can well disambiguate entities for French textual content. As future work, we aim to create a bigger evaluation dataset and to share it with the NLP community. We are also interested in finding a way to detect if no candidate entity matches, allowing us to spot missing data in the network. To do so, we plan to use the Deep Semantic Similarity Measure (DSSM) method (Huang *et al.*, 2013) in order to directly compute similarities between the nodes and identify if the entity belongs to the network or not. Finally, to solve the word2vec problem, we plan to use the FastText library (Bojanowski *et al.*, 2016) since it is robust against unknown vocabulary words because it uses letter-grams instead of tokens to compute embeddings. We can also improve the model by adding the full Wikipedia dump into the frWac corpus training data.

Références

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

DAIBER J., JAKOB M., HOKAMP C. & MENDES P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *9th International Conference on Semantic Systems (I-SEMANTICS)*, Graz, Austria.

DBPEDIA (2007). Dbpedia. `http://wiki.dbpedia.org`.

FRANZONI V., MENCACCI M., MENGONI P. & MILANI A. (2014). Heuristics for semantic path search in wikipedia. In *14th International Conference on Computational Science and Its Applications (ICCSA)*.

FREEBASE (2007). Freebase. `https://www.freebase.com`.

GEONAMES (2006). Geonames. `http://www.geonames.org`.

GRAVIER G., ADDA G., PAULSSON N., CARR M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *8th International Conference on Language Resources and Evaluation (LREC)*.

HACHEY B., NOTHMAN J. & RADFORD W. (2014). Cheap and easy entity evaluation. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

HUANG P.-S., HE X., GAO J., DENG L., ACERO A. & HECK L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information & Knowledge Management (CIKM)*.

LAFOURCADE M. (2007a). Jeux de mots. `www.jeuxdemots.org/`.

LAFOURCADE M. (2007b). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *7th International Symposium on Natural Language Processing (SNLP'07)*.

LIBEN-NOWELL D. & KLEINBERG J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*.

LINKEDMDB (2009). Linkedmdb. `http://www.linkedmdb.org`.

LÜ L., JIN C.-H. & ZHOU T. (2009). Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E*.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a french lexical network : Methodological issues.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*.

MORO A., RAGANATO A. & NAVIGLI R. (2014). Entity linking meets word sense disambiguation : a unified approach. *TACL*.

MUSICBRAINZ (2000). Musicbrainz. `https://musicbrainz.org`.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*.

NUZZOLESE A., GENTILE A., PRESUTTI V., GANGEMI A., GARIGLIOTTI D. & NAVIGLI R. (2015). The 1st Open Knowledge Extraction Challenge. In *12^{th} European Semantic Web Conference (ESWC)*.

PLU J. (2016). Knowledge Extraction in Web Media : At The Frontier of NLP, Machine Learning and Semantics. In *25^{th} World Wide Web Conference (WWW), PhD Symposium*.

SAGOT B. & FIER D. (2008). Building a free french wordnet from multilingual resources. In *Ontolex 2008*.

SONG M., HEO G. & DING Y. (2015). Sempathfinder : Semantic path analysis for discovering publicly unknown knowledge. *Journal of Informetrics*.

STERN R., SAGOT B. & BÉCHET F. (2012). A joint named entity recognition and entity linking system. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*.

WIKIDATA (2012). Wikidata. `https://www.wikidata.org`.