An Experimental Study of a Hybrid Entity Recognition and Linking System

Julien Plu, Giuseppe Rizzo, Raphaël Troncy

EURECOM, Sophia Antipolis, France
{julien.plu,giuseppe.rizzo,raphael.troncy}@eurecom.fr

Abstract. We present an experimental study of the performance of a hybrid semantic and linguistic system for recognizing and linking entities from formal and informal texts. In the current literature, systems are generally tailored to one or a few types of textual documents (e.g. narrative texts, newswire articles, informal text such as microposts). In contrast, we assess the performance of a hybrid approach that adapts the entity extraction, recognition and linking process to the type of document being analyzed. The hybrid system relies on POS taggers, gazetteers and Twitter user account dereferencing modules to extract entities, NER modules to recognize and type entities and entity popularity, string distance measures and scoring functions to disambiguate entities by ranking potential link candidates. The evaluation results show the robustness of our proposed approach in terms of text-independence compared to the current state-of-the art.

Keywords: Entity Extraction, Entity Linking, Entity Recognition, Entity Filtering

1 Introduction

The approach we analyze in this paper links entities from both formal and informal textual documents to DBpedia 2014 resources. The approach can be broken down into three tasks: entity extraction, entity recognition and entity linking. Entity extraction refers to the task of spotting mentions that can be entities in the text. Entity recognition refers to the task of giving a type to the extracted entity. Entity linking refers to the task of linking the mention to a targeted knowledge base, and it is often composed of two sub-tasks: generating candidates and ranking them according to scoring functions. We experiment with three standard benchmark corpora: two composed of tweets and one composed of formal texts.

Numerous approaches have been proposed to address the task of extracting, typing and linking entities in formal (such as newswire content) and informal (such as microposts) text. Among the recent and best performing systems, both WAT [8] and TagME [7] turn the task in a multi-stage pipeline composed of extraction, typing, linking and pruning. The two systems have been tested on both types of text. AIDA [3], Babelfy [6] and DBpedia Spotlight [4] are other systems that well-perform when processing formal text. Specific methods have been developed for analyzing tweets, such as the E2E system [5], which has introduced a new methodology for the entity linking: they recast the extraction and linking steps as a single process. Two other similar approaches are AIDA for tweets [1], and DataTXT [11] (extension of TagME [7]), which have adapted their original well-performing approach on formal text to tweets.

2 Architecture

Our goal is to link all the entities occurring in a text to their counterparts in DBpedia 2014. Entities that do not have an entry in the knowledge base will be linked to *NIL*. Tweets are notoriously problematic to process compared to formal texts because of *i)* hashtags (such as *#barackobama* referring to *Barack Obama*); *ii)* user mentions (e.g. *@ryeong9* referring to *db:Ryeo_Wook*[1]); *iii)* acronyms (e.g. *Met* for *Metropolitan Police Service*); *iv)* short length of only 140 characters and *v)* syntax which is often grammar free and words are misspelled. In the following, we briefly describe the different components of our hybrid approach.

Text Normalization. Only for tweets, this step consists of normalizing the text. We remove emoticons, extra white spaces and punctuation symbols belonging to two unicode categories[2]: other and symbol.

Entity Extraction and Recognition. This step is about detecting mentions from text that are likely to be selected as entities. The objective is achieved by using the following components: 1) POS tagger; 2) NER; 3) gazetteer and 4) time expressions spotter. For tweets, we add a dereferencing Twitter account system to retrieve the user name. Our POS tagging system is the Stanford NLP POS Tagger with a model trained specifically for tagging tweets[3] in order to be case insensitive and to get independent tags for mentions and hashtags. For formal text we use the *english-bidirectional-distsim* that provides a better precision but for a higher computing time. With the POS tagger, we spot all the noun phrases and numbers. As NER system, we use Stanford NER properly trained with the training set of a benchmark. The gazetteers reinforce this stage bringing a robust spotting for well-known nouns. Tweets contain Twitter accounts, so we dereference Twitter accounts and extracts their public names. Each of those systems are launched in parallel. The type of an entity is given by Stanford NER. If a mention is not detected by Stanford NER then no type is assigned.

Entity Resolution. Given two overlapping mentions, e.g. *States of America* from Stanford NER and *United States* from Stanford POS tagger, we only take the union of the two phrases. We obtain the mention *United States of America* and the type provided by Stanford NER is selected.

Entity Linking. This step is composed of three sub-tasks: *i)* entity generation, where we lookup up the entity in an index built on top of both DBpedia2014[4] and a dump of the Wikipedia articles[5] dated from October 2014 to get possible candidates; *ii*) candidates filtering based on direct inbound and outbound links between the extracted entities in Wikipedia; *iii)* entity ranking based on an in-house ranking function using Levenshtein distance between the extracted mention and the title, the set of redirect pages and the set of disambiguation pages of each candidate, weighted by their PageRank. If an entity does not have an entry in the knowledge base, we normally link it to

[1] db stands for `http://dbpedia.org/resource/`

[2] `http://www.fileformat.info/info/unicode/category/index.htm`

[3] `https://gate.ac.uk/wiki/twitter-postagger.html`

[4] `http://wiki.dbpedia.org/services-resources/datasets/datasets2014`

[5] `https://dumps.wikimedia.org/enwiki/`

NIL. The detailed method followed for the linking step and how to build the index is explained in [9].

3 Experimental Study

Our hybrid approach has been tested against the test dataset of the #Micropost2014 [2] and #Micropost2015 [10] NEEL challenges and the OKE2015 challenge[6]. The breakdown results for each of these datasets are available[7]. Table 1 shows the performance of our approach in comparison with state-of-the art systems given the F1-measure at the final linking stage.

	Our Approach	DBPedia Spotlight	TagME (DataTXT)	Babelfy	WAT	AIDA	E2E	UTwente	ousia	acubelab
#Microposts2014	46.29	*N/A*	49.9	*N/A*	*N/A*	45.37	**70.06**	54.93	*N/A*	*N/A*
#Microposts2015	47.95	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	**80.67**	47.57
OKE2015	**48**	39.8	37.9	42	*NA*	44.6	*N/A*	*N/A*	*N/A*	*N/A*

Table 1. F1-measure results at the linking stage on both the #Microposts2014 NEEL challenge and OKE2015 challenge test datasets

Results for #Micropost2014 and #Micropost2015 NEEL challenges come from the official published results. We report on the best performing systems: E2E, DataTXT, AIDA, UTwente for #Micropost2014 and ousia, acubelab for #Micropost2015. For the OKE2015 challenge, we have changed the test dataset in order to fix annotation issues, those changes has been approved by the organizers and committed to the official repository. We used the neleval[8] scorer instead of the official one used in the challenge. This explains why the results are different than the ones reported by the organizers of the challenge. Those datasets have differences listed in Table 2. We have chosen those different datasets to show the full potential of our approach which can be adapted depending on the dataset to be processed in terms of features and kind of text.

Datasets	co-references	typing	NIL entities	dates	numbers
OKE2015	✓	✓	✓	✗	✗
#Microposts2014	✗	✗	✗	✓	✓
#Microposts2015	✗	✓	✓	✗	✗

Table 2. Features for each datasets

The results that are reported for DBpedia Spotlight, TagME, AIDA and Babelfy for OKE2015 have been obtained using their respective APIs with the best tested settings. For DBpedia Spotlight, those settings are: *confidence=0.3* and *support=20*. For TagME, those settings are: *include_all_spots=yes* and *epsilon=0.5*. For AIDA those settings are: *technique=GRAPH*, *algorithm=COCKTAIL_PARTY_SIZE_CONSTRAINED*, *alpha=0.6* and *coherence=0.9*. For Babelfy, those settings are: *lang=en*, *annType=NAMED_ENTITIES*, *annRes=WIKI*, *match=EXACT_MATCHING*, *dens=true* and *th=0.4*. Since WAT is not publicly accessible, we did not test it with the OKE challenge dataset. For formal text (OKE benchmark), our system outperforms the other approaches

[6] https://github.com/anuzzolese/oke-challenge
[7] http://multimediasemantics.github.io/adel/
[8] https://github.com/wikilinks/neleval

being tested. For informal text (NEEL corpora), the hybrid approach shows the robustness in extracting and typing entities because we jointly use linguistic and semantic methods. The results slightly drop at linking level.

4 Conclusion and Future Work

In this paper, we have presented the experimental study of a text independent hybrid approach that extracts, recognizes and links entities to DBpedia 2014. We show that a successful approach exploits both linguistic features and semantic features extracted from DBpedia. As future work, we plan to focus on improving the linking task by making more use of graph-based algorithms, and to improve our ranking function.

Acknowledgments

This work was partially supported by the innovation activity 3cixty (14523) of EIT Digital (`https://www.eitdigital.eu`).

References

1. Y. M. Amir, H. Johannes, I. Yusra, B. Artem, and W. Gerhard. Adapting AIDA for Tweets. *4^{th} International Workshop on Making Sense of Microposts (# Microposts2014)*, 2014.
2. A. E. Cano, G. Rizzo, A. Varga, M. Rowe, Stankovic Milan, and A.-S. Dadzie. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *4^{th} International Workshop on Making Sense of Microposts (#Microposts2014)*, 2014.
3. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *8^{th} Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.
4. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7^{th} International Conference on Semantic Systems (I-Semantics)*, 2011.
5. C. Ming-Wei, H. Bo-June, M. Hao, L. Ricky, and W. Kuansan. E2E: An End-to-End Entity Linking System for Short and Noisy Text. In *4^{th} International Workshop on Making Sense of Microposts (# Microposts2014)*, 2014.
6. A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014.
7. F. Paolo and S. Ugo. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *19^{th} International Conference on Information and Knowledge Management (CIKM)*, 2010.
8. F. Piccinno and P. Ferragina. From TagME to WAT: a new entity annotator. In *1^{st} ACM International Workshop on Entity Recognition & Disambiguation (ERD)*, 2014.
9. J. Plu, G. Rizzo, and R. Troncy. A Hybrid Approach for Entity Recognition and Linking. In *12^{th} European Semantic Web Conference, Open Knowledge Extraction Challenge*, 2015.
10. G. Rizzo, Cano Amparo E, B. Pereira, and A. Varga. Making sense of Microposts (#Microposts2015) named entity recognition & linking challenge. In *5^th International Workshop on Making Sense of Microposts (#Microposts'15)*, 2015.
11. S. Ugo, B. Michele, P. Stefano, P. Gaetano, D. T. Emilio, and V. Mario. DataTXT at# Microposts2014 Challenge. *4^{th} International Workshop on Making Sense of Microposts (# Microposts2014)*, 2014.