Graph-based named entity identification

Sammy Khalife, Michalis Vazirgiannis
Ecole Polytechnique, LIX, Dascim
khalife@lix.polytechnique.fr

Abstract

Named entity discovery (NED) is an important research problem that can be decomposed into two sub-problems. Firstly, named entity recognition (NER) to extract or tag pre-defined sets of words in a vocabulary (called "named entities": names, places, locations, ...) when they appear in natural language text. Second, named entity linking/identification (NEL) considers these mentions as queries to be identified in a pre-existing database.

In this work, we consider the Named entity linking problem, and assuming a set of queries (or mentions) that have to be identified within a knowledge base, represented by a text database paired with a semantic graph. We present state-of-the-art methods in NEL, and propose a 2-step method for the identification of named entities.

First, a filtering method, capitalizing on text and graph mining techniques, aiming to maximize recall at K (typically for $5 \leq K \leq 10$). Second, a scoring scheme is introduce to maximize precision at 1 by re-ranking the remaining top candidates using the query (or mention), enriched type features, and graph information. We present our algorithms in details, and show on NIST TAC-KBP datasets (http://www.nist.gov/tac/) they reveal to be efficient compared to previous approaches.

Contribution

Our work is aiming to provide an overview of the NEL problem, to propose a 2-step method based on entity filtering and graph mining, and to compare its performance to other approaches and show its advantages. The approach we propose is well motivated from potential industrial applications. Social networks, or industrial compliance processes are natural applications of named entity identification. In such cases A NEL system can be useful for semi-automatic document comprehension.

1 Related work

Name entity linking implies having a database at hand. **Collective linking** set the correspondances jointly with all mentions and entities. In **individual linking**, queries are supposed indenpendant. For more details on related work, references are given below.

1.1 Main references

- **Graphs for NED** : Dense subgraph extraction [6], [2] , Hubs and authorities : HITS algorithm [9]
- **Graphical models** Modified HMMs for NED [1],
  Probabilistic bag of Hyperlinks [4], Evidence mining for NED [7]
- **Deep learning architectures** Learning embeddings for NED [10], Convolutional networks [8]

1.2 Comments

Each of these approaches uses a filtering metric to discard non relevant entities : the final precision score will be upper-bounded by the recall of the filtering.

We did not consider deep learning related algorithms in our experiments for the following reasons :

- Difficult to obtain a large datasets of annotated samples
- Data quality of automatic generated mentions from Wikipedia is debatable
- Neural networks require large datasets

In our work, only **individual linking** is considered.

2 Graph-based ranking algorithm

In this section, we propose a graph mining procedure for *named entity identifcation* to extract semantic information that filtering cannot capture. Our 2-step method is as follows :

- Filtering method (cf next subsection) to maximize recall at $\boldsymbol{K}$ ($\boldsymbol{R}$@$\boldsymbol{K}$, $5 \leq \boldsymbol{K} \leq 10$)
- Enriched features extraction from the knowledge graph
- Supervised re-ranking the $\boldsymbol{K}$ top entity candidates

2.1 Filtering method

Every state-of-the-art solution proposed uses a filtering procedure to discard an important majority of entities. We propose a mention name scoring using acronym detection, context TF-IDF scoring to improve previous filtering methods ([3] for acronym detection and [12] for an application to entity linking).

2.2 Re-ranking with new ontology types

2.2.1 Type matching

To add semantic information, we attribute for each entity type, a corresponding set of entity types. Let T the family of ontology types. Such mapping is defined as follows

$$\begin{aligned} \varphi \in \Phi : T &\longrightarrow \{0,1\}^{|T|} \\ \mathbf{t} &\longmapsto (\tilde{t}_1, ..., \tilde{t}_n) \end{aligned}$$

We aim to find the (or a) best type mapping $\hat{\phi}$ defined on all ontology types, providing an optimal set of new entities. Let γ a positive function defined on $\mathbb{R}_+^{|T|}$ (for example, mean, max or min function). We define γ_ϕ a functon taking as input the vector of aggregated scores :

$$\gamma_\phi(q_i, e_j) = \gamma(s(q_i, e_j), s(q_i, n_1^j), ..., s(q_i, n_{|T|}^j)$$

If $deg(e_j) < |T|$, then we complete scores by 0 or use graph exploration to complete the score vector (cf comments subsection)

H a loss function (for example, $x \longmapsto \max(0, 1-x)$). For a query indexed by i, let $\mathcal{C}_i$ the set of candidate entities after filtering, except the gold entity. An empirical optimal mapping $\hat{\phi}$ is defined as follows :

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i} H(\gamma_\phi(q_i, e_{G_i}) - \gamma_\phi(q_i, e_j)) \tag{1}$$

Using a boolean formulation fort ϕ, (1) can be interpreted as combinatorial optimization problem. We considered simple type mappings first. To do so, we :

- Displayed mis-ranking by entity types
- We constraint ϕ to a subset of mapping functions. To do so, we extracted new graph features only for a set most frequent ontology types concerned by identification-errors [for ex., cf figure 1].

2.2.2 Example : Cities

T represents the types "City", "State", "Artist", "Museum", "Country", "FootballPlayer". Example of type mapping function for "City" type is $T[0] = (1, 1, 0, 1, 1, 0)$: only country, states and related museums are considered to identify a city [cf figure 1].

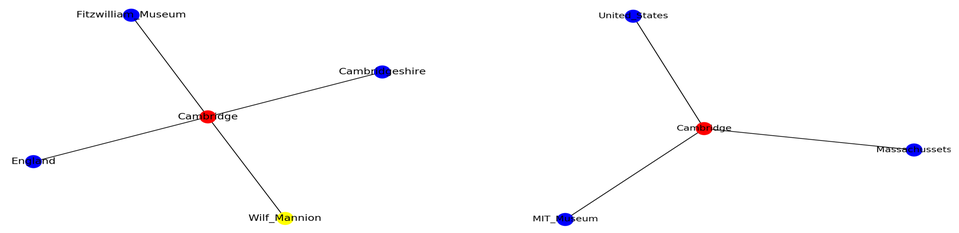

Figure 1: Two homonyms : Cambridge cities. To the left, in England. To the right, in Massachussets, in the United States. A neighbor entity e is in blue if $\phi(T_e) = 1$ and yellow otherwise (Wilf Mannion is a former Cambridge football Player in England).

2.2.3 Graph mining procedure : comments

Based on the previous statements and Figure 1, we bring forward a proposal the following scheme :

- Keep an explicit list of informative entities $(t_i)_{1 \leq i \leq T}$ and constant mapping function ϕ such that $\phi(\tilde{t})[t_i] = 1$ and 0 otherwise for any type $\tilde{t}$. For example $(t_i)_{1 \leq i \leq T}$ can correspond to countries, states, and towns.
- Using ϕ, build a score vector using neighborhood of entity candidates. This score vector can be directly computed with TF-IDF
- Concatenate new scores into a features vector (cf algorithm 1)

Algorithm 1 Knowledge graph mining

Require: Knowledge Graph $\boldsymbol{G}$, Query $\boldsymbol{Q}$, Entity top candidates with initial filtering score $(\boldsymbol{E}_i, \boldsymbol{S}_0^i)_{1 \leq i \leq K}$, Projection types $(\boldsymbol{T})_{1 \leq j \leq |T|}$, Disambiguation map $\boldsymbol{\phi}$

1: **for** $i = 1$ to K **do**
2: Get neighbors or close nodes in the graph $(\boldsymbol{N}_j(\boldsymbol{E_i}))_{1 \leq j \leq |T|}$ in G according the the type mapping ϕ
3: Build scores vector $\{\boldsymbol{S}_0^i\} \cup \{\boldsymbol{S}^i_{\boldsymbol{N}_j 1 \leq j \leq |T|}\}$ for each type node using scoring TF-IDF comparison between entity description and the query $\boldsymbol{Q}$
4: **end for**
5: **return** Score vectors $(\boldsymbol{S}^i)_{1 \leq i \leq K}$

3 Experiments : datasets and results

Dataset	Non-NIL mentions			Total
	PER	ORG	GPE	
2009 (Test)	255	1013	407	1675
2010 (Test)	213	304	503	1020
Total (Test)	468	1317	910	2695
2010 (Train)	335	335	404	1074
2014 (Train)	1461	767	1313	3541
Total train	1796	1102	1717	4615

Figure 2: Number of non-nil mentions in NIST TAC-KBP Datasets

	R@100		R@10		P@1	
	TAC09	TAC10	TAC09	TAC10	TAC09	TAC10
[8]	(*)	(*)	(*)	(*)	82.26	83.92
[11]	/	94.6	/	/	/	87.7
[10]	/	/	/	/	/	85.2
[5]	/	/	/	/	/	87.2
M1	98.93	99.61	97.73	98.73	88.54	87.06
M2	98.93	99.61	97.73	98.73	**89.85**	**88.82**

Figure 3: Comparison of our apporach to state-of-the art methods

M1 : entity filtering. M2 : filtering and graph mining re-ranking with supervised learning (logistic regression) State-of-the-art precision and recall on NIST TAC-KBP 2009/2010 datasets [Needs to add other datasets for statistical significance]

Conclusions

In this work, we provided a survey of existing algorithms for named entity identification and showed that such identification can be improved using adapted text and graph mining techniques. Our 2-step routine uses a filtering algorithm based on information retrieval techniques and acronym expansion. Then, remaining entity candidates features are mapped with semantic scores from the knowledge graph using a type mapping function.

Advantages of our method over deep learning : First, entity features are interpretable and our method does not require lots of data. Finally, this graph-based routine leads to potential future improvements: first, by optimizing mapping ϕ with meta-heuristics (such as genetic algorithm), and second by compare local-graph structure to enrich entity features (using graph kernels for instance).

Acknowledgements

We would like to thank all Dascim LIX team members for their help, especially Christos Giatsidis concerning the use of Hadoop and Spark technologies. Also, thanks go to Octavian Ganea (ETH Zurich) for his time and constructive discussions. This work was partially supported by by Association Nationale de la recherche et de la technologie (ANRT) and BNP Paribas under the Cifre convention 2016/2017.

References

[1] Ayman Alhelbawy and Robert Gaizauskas. Named entity disambiguation using hmms. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03*, pages 159–162. IEEE Computer Society, 2013.

[2] Ayman Alhelbawy and Robert J Gaizauskas. Graph ranking for collective named entity disambiguation. In *ACL (2)*, pages 75–80, 2014.

[3] Maud Ehrmann, Leonida Della Rocca, Ralf Steinberger, and Hristo Tanev. Acronym recognition and processing in 22 languages. *arXiv preprint arXiv:1309.6185*, 2013.

[4] Octavian-Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. Probabilistic bag-of-hyperlinks model for entity linking. In *Proceedings of the 25th International Conference on World Wide Web*, pages 927–938. International World Wide Web Conferences Steering Committee, 2016.

[5] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 621–631, 2016.

[6] 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*, pages 782–792. Association for Computational Linguistics, 2011.

[7] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1070–1078, New York, NY, USA, 2013. ACM.

[8] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*, pages 1333–1339, 2015.

[9] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. Agdistis-graph-based disambiguation of named entities using linked data. In *International Semantic Web Conference*, pages 457–471. Springer, 2014.

[10] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*, 2016.

[11] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Learning distributed representations of texts and entities from knowledge base. *arXiv preprint arXiv:1705.02494*, 2017.

[12] Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. Entity linking with effective acronym expansion, instance selection, and topic modeling. In *IJCAI*, volume 2011, pages 1909–1914, 2011.