Entity Linking with Multiple Knowledge Bases: an Ontology Modularization Approach

Bianca Pereira

Insight Centre for Data Analytics
National University of Ireland, Galway
bianca.pereira@insight-centre.org

Abstract. The recognition of entities in text is the basis for a series of applications. Synonymy and Ambiguity are among the biggest challenges in identifying such entities. Both challenges are addressed by Entity Linking, the task of grounding entity mentions in textual documents to Knowledge Base entries. Entity Linking has been based in the use of single cross-domain Knowledge Bases as source for entities. This PhD research proposes the use of multiple Knowledge Bases for Entity Linking as a way to increase the number of entities recognized in text. The problem of Entity Linking with Multiple Knowledge Bases is addressed by using textual and Knowledge Base features as contexts for Entity Linking, Ontology Modularization to select the most relevant subset of entity entries, and Collective Inference to decide the most suitable entity entry to link with each mention.

Keywords: Entity Linking, Linked Data, Ontology Modularization

1 Problem Statement

Natural language understanding, in particular, the recognition of entities in text, has been the basis for different computer-based applications such as Semantic Search, Recommendation Systems, Sentiment Analysis, and Social Media Monitoring just to mention a few.

Among the biggest challenges in the recognition of entities are the synonymy and ambiguity. Synonymy is given by the existence of different names (mentions) in text to the same real world entity. For instance, *IBM* and *International Business Machines* are two mentions for the same real world entity. Ambiguity, moreover, occurs when one mention may refer to more than one real world entity. The mention *Jackson*, for instance, may refer to more than 500 different entities [1].

Entity Linking is the task of *grounding entity mentions in documents to Knowledge Base entries* [7]. It is investigated as way to solve both synonymy and ambiguity by linking mentions in natural language text with entity entries in a given Knowledge Base, where those entries provide a unique identifier for each

[1] http://en.wikipedia.org/wiki/Jackson_(name)

real world entity. One example is the interlinking between the mention *Michael Jackson* and the Wikipedia [2] link `http://en.wikipedia.org/wiki/Michael_Jackson` providing information about which real world entity the mention is referring to.

Cross-domain datasets such as Wikipedia, DBPedia [3] and YAGO [4] are the most used Knowledge Bases for Entity Linking. This happens mainly because of the high number of domains they cover. Even so, those Knowledge Bases can benefit from expanding them using other Knowledge Bases. Wikipedia, for instance, has data about almost 5 million entities in different domains. Internet Movie Database [5] (IMDB) contains more than 8 million entities in the cinema domain, and Music Brainz [6] has 30 million entities in the music domain. Those two databases contain seven times more entities than whole Wikipedia even when only two domains are considered.

The focus of this work is to enrich the understanding of entities in natural language text through Entity Linking using multiple Knowledge Bases.

2 Relevancy

Entity Linking enables the improvement of a broad number of applications, such as: the enrichment of the reader experience when applied in Wikification; the query for documents based on real world entities instead of keywords, when applied to Semantic Search; the recognition of sentiments related to entities when applied to Sentiment Analysis or Reputation Management; and the enrichment of the analysis of textual data in Big Data Analytics Solutions. The recognition of entities through Entity Linking can also be used for a series of Information Extraction (e.g. Relation Extraction and Attribute Extraction) and Natural Language Processing approaches (e.g. Coreference Resolution across texts).

The use of multiple Knowledge Bases for Entity Linking improves even more those applications. It increases the range of entities they can work with. The combination of private and public Knowledge Bases can even be applied to understand enterprise textual data (such as reports, intranet texts, and Customer Relationship Management descriptions) improving the value of the data for business.

3 Related Work

There are relatively few work on using multiple Knowledge Bases for Entity Linking. Tools such as AlchemyAPI [7] and Zemanta [8] are two examples but due

[2] http://www.wikipedia.org/
[3] http://dbpedia.org/
[4] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[5] http://www.imdb.com
[6] http://musicbrainz.org
[7] http://www.alchemyapi.com/
[8] http://www.zemanta.com/

to their commercial characteristic, the methods they use are not described in detail. In the research side, there is the work from [2] and [12] dealing with more than one Knowledge Base for Entity Linking.

[12] performs Entity Linking using Relational Databases. The authors present an adaptive solution that can use any database under Boyce-Codd Normal Form (BCNF) as Knowledge Base. They used IMDB and a sports database for evaluation of their approach.

[2] focuses on large scale Entity Linking using Linked Data. DBPedia, Freebase [9], Geonames [10] and New York Times Linked Data [11] are used as Knowledge Bases. The authors assume all Linked Data datasets are already linked to each other. Their algorithm relies on the existence of textual descriptions in the Knowledge Base and in a crowd-sourcing step to improve the results.

[2] uses TF-IDF over the textual descriptions in the Knowledge Base to choose the best entity to be linked with each mention. [12] uses text in entity attributes in the Knowledge Base as information for disambiguation. Both solutions use only the words around each mention as context for disambiguation (*local context*). Despite the state-of-the-art in Entity Linking shows the improvement in using more information as context for disambiguation [11], those contexts (*local and global*) were not applied for Entity Linking with Multiple Knowledge Bases.

State-of-the-art solutions describe the number of candidates for each mention (i.e. the ambiguity of a mention) as the main drawback for Entity Linking [4, 3, 6]. In other words, as bigger the Knowledge Base and the number of candidates per mention as much time it may take to find a solution. While current approaches rely in approximate algorithms, this PhD research proposal aims the use of Ontology Modularization to limit the number of entity entries considered for linking. Ontology Modularization has been used to improve manual annotation of image and text [1, 13]. To the best of our knowledge this is the first time Ontology Modularization is used for Entity Linking.

4 Research Questions

The problem of Entity Linking using multiple Knowledge Bases leads to the following research questions:

RQ1 What are the textual features most relevant for Entity Linking?

RQ2 What are the Knowledge Base features most relevant for Entity Linking?

RQ3 Is it feasible to create an Entity Linking approach using multiple Knowledge Bases with comparable performance to Knowledge Base-specific approaches?

[9] http://www.freebase.com/
[10] http://www.geonames.org/
[11] http://data.nytimes.com/

5 Hypotheses

Our work is based on several hypotheses related to the research questions listed in Section 4:

H1 Verbs and mentions appearing near a given mention measures the relevance of a candidate entry to be linked with this mention. (local context)
H2 Mentions appearing in the same paragraph are more relevant for disambiguation than those appearing far in the text. (global context)
H3 Mentions and verbs in text can be directly mapped to entities and relationships in the Knowledge Base.
H4 Comparable performance with Knowledge Base-specific Entity Linking can be achieve through the division of the Knowledge Base in context-specific modules.

6 Preliminary Results

Our first experiments envisioned the creation of a baseline approach for Entity Linking with Multiple Knowledge Bases.

The first experiment analyzes the feasibility on using solely Linked Data Knowledge Bases as sources for entity mentions in text. This experiment evaluates different heuristics to automatically discover properties that carry lexical information for entity entries in the Knowledge Base. The results are available in [10] and they shows that even simple heuristics can identify the correct properties. It demonstrates that Linked Data Knowledge Bases can be used as dictionary for mentions.

The second experiment verifies if current linking methods developed for a specific Knowledge Base can be used with different ones. We adapted the work presented in [3] to use Jamendo [12] and Linked Movie Database [13], two publicly available Linked Data datasets. As those datasets do not have full text descriptions for all entities, the mapping between text around mentions and textual descriptions could not be used, as well as the computation of TF-IDF. Even without this context the results were quite encouraging with f-score of 54% for Disambiguation with Jamendo and 87% with Linked MDB. Those results are presented in [9].

Due to the positive results from the second experiment, the third one envisions the use of DBPedia in order to compare with related work. Using AIDA-CoNLL[4], an annotated corpora, as gold standard, the best accuracy was 32% while related work [8] reports accuracy ranging from 34% to 82.3% in the same corpora. Thus, the next step is to evaluate how much each textual and Knowledge Base features contribute to find the best linking. This analysis will help us discover why the results with DBPedia were so low when they were good with the other Knowledge Bases. More information about future evaluations are given in Section 8.

[12] http://dbtune.org/jamendo/
[13] http://linkedmdb.org/

7 Approach

Natural language text often contains a human-readable description of a set of entities and their relationships. We consider Knowledge Base as a machine-readable description of entities and their relationships. This approach aims to use a model that better represents both descriptions in order to identify the best pair (mention, Knowledge Base entry) that refers to the same real world entity.

There are many formats of Knowledge Bases that can be used for Entity Linking. Linked Data datasets explicit the semantic of the data and enable the easy extension of the Knowledge Base by direct linking entries in different datasets. Because of this, this work focus only on the use of Linked Data datasets as Knowledge Bases.

A series of constraints were identified during the development of our baseline:

- There is no annotated corpora associated to the Knowledge Base.
- The Knowledge Base does not contain textual description for entities.
- All information available for Entity Linking comes only from the textual source and the Knowledge Base.

This approach is divided in three sequential steps (Figure 1). The Mention Recognition step deals with the identification of entity mentions in text. The Candidate Selection step collects a list of candidate entity entries in the Knowledge Base to link with each mention. And the Disambiguation step selects the most suitable candidate to be linked with each mention. The solution applied for each one of these steps is explained in the following subsections.

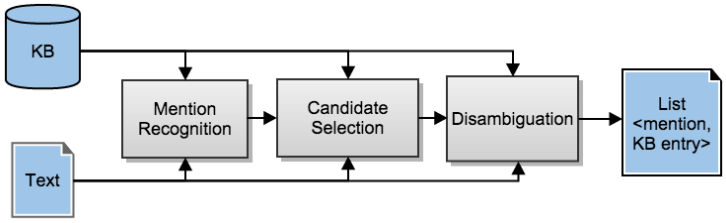

Fig. 1. Entity Linking pipeline.

7.1 Mention Recognition

Mention Recognition is the first step for Entity Linking. This step deals with the recognition of entity mentions in text. The text is given as input and the expected output is a list with all mentions appearing in it. Using the text *"Mick Jackson wrote Blame it on the Boogie, a great success of The Jackson's."* as input text, the expected output is the list M = {"Mick Jackson", "Blame it on the Boogie", "The Jackson's"}.

Each entry in the Knowledge Base is considered as an entity entry if there is lexical information attached to it. In other words, a given entry is considered an entity entry only if its name is expressed in the Knowledge Base. Based on this,

the Knowledge Base is used as a dictionary to identify mentions in the text. If one surface form in the text matches a dictionary entry then the surface form is recognized as a mention.

7.2 Candidate Selection

Given all mentions recognized by the Mention Recognition step, the Candidate Selection step selects a set of candidate entries for each mention. In this context, candidate entry is a Knowledge Base entry with a probability higher than zero to be linked with a given mention. The biggest issue in candidate selection is the high number of possible candidates for a single mention. For instance, the word *Jackson* may refer to more than 500 entity entries in Wikipedia [14]. Considering a Knowledge Base with all people in the world, the number of candidates for the word Jackson turns the process of Linking unfeasible. Due to that, the Candidate Selection should return only the most relevant subset of candidates for the given text.

The contribution of this work in the candidate selection is in the use of Ontology Modularization. The process of Candidate Selection is given by the division of the Knowledge Base in contextual modules. Each contextual module keeps a maximum limit on the number of candidates for each surface form. The key idea is that entities in the same context are semantically related in the Knowledge Base. Thus, by looking for candidates in the Knowledge Base, the module that contains candidates for the highest number of mentions is more likely to represent the context of the text.

7.3 Disambiguation

Finally, the Disambiguation step chooses the most suitable entity entry to be linked with each mention. If there is no suitable entry for a given mention this step should return NIL instead. The main challenge in disambiguation is identifying the best entity entry by using the textual content in the document and the information provided by the Knowledge Base.

Considering the following text: *"The Battle of the Boogie was the event where Michael Jackson, the writer of the music, and Michael Jackson, from The Jackson 5, were in a dispute for chart positions."*. A human reader easily understands who are the two Michael Jacksons in the text. By that, she uses hers background knowledge and the context given by the words in the text to infer the difference between both. The aim of this approach is to have an algorithm that mimic this process. It collects the textual context (Research Question RQ1 and Hypotheses H1 and H2) and maps it to the context provided by the Knowledge Base (Research Question RQ2 and Hypothesis H3). A Collective Inference model will be used to merge both contexts and identify the best (mention,entry) link [5].

The contribution of this work is in using only content words near mentions (Hypothesis 1) and lexical cohesion in the text (Hypothesis 2) as textual contexts. And the mapping between those content words (nouns, adjectives, and verbs) to semantic information in the Knowledge Base (entities, attributes, and relationships) rather than to textual descriptions.

[14] http://en.wikipedia.org/wiki/Jackson_(name)

8 Evaluation Plan

There are some corpora already linked to Wikipedia and Wikipedia-related Knowledge Bases [4, 11]. Those corpora can be used as gold standard for Entity Linking and they enable the comparison with state-of-the-art approaches [8]. The unique problem is that such corpora do not enable the evaluation using other Linked Data datasets as Knowledge Bases. Given that, the first step for the evaluation will be the development of a set of Guidelines for Annotation and the annotation of a corpus with links to different Knowledge Bases.

8.1 Mention Recognition

In the Mention Recognition step, the evaluation will measure the coverage of the Knowledge Base as a dictionary to recognize mentions. This evaluation will use a manually annotated corpora as gold standard and precision, recall, and f-measure as measures for evaluation.

8.2 Candidate Selection

The best method for generation of modules is the one which returns all the best possible candidates as part of the candidate set. The annotated corpora will be used as gold standard for evaluation and the accuracy will be used as the measure to evaluate the Ontology Modularization method.

8.3 Disambiguation

For disambiguation it is necessary to evaluate: the contribution of each textual feature, the contribution of each Knowledge Base feature, and the contribution of the methods for collective inference using both textual and Knowledge Base contexts. The first two contributions can be measured separately by generating a ranking of candidates for each mention in text and measuring the precision@N, recall@N, and f-measure@N. In other words, measuring if the best entity entry given by the gold standard appears in the first top N positions in the ranking. This separation of each feature enables the comparison with features used in related work. The evaluation of methods for collective inference will be made by choosing the best features from text and Knowledge Base and measuring the results using precision, recall and f-measure. It also enables the comparison with related work.

9 Reflections

Despite the increasing number of research papers in Entity Linking there are only a small number of works considering the use of more than a single Knowledge Base. In this PhD research proposal the goal is to solve the problem of Entity Linking with Multiple Knowledge Bases. This will be done by using different textual and Knowledge Base features, and Ontology Modularization to select entities in the same semantic context. The main contributions are in enabling: a higher number of Knowledge Bases to be used for Entity Linking, the use of

Linked Enterprise Data in Entity Linking context, and the use of more types of entities for Entity Linking and applications based on the use of entities.

In conclusion, this work will improve upon previous approaches by enabling the use of multiple Knowledge Bases for Entity Linking.

Acknowledgments.. The author would like to thanks Georgeta Bordea for the kind review of early drafts of this paper as well as Emir Munoz and Andre Freitas for the insightful discussions. This work has been conducted under the supervision of Paul Buitelaar and was partially funded by the EC for the FP7 project EuroSentiment under Grant Agreement 296277 and in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 for the INSIGHT project.

References

1. Mathieu D'Aquin, Anne Schlicht, Heiner Stuckenschmidt, and Marta Sabou. Ontology Modularization for Knowledge Selection: Experiments and Evaluation. In *Database and Expert Systems Applications. LNCS*, pages 874–883. 2007.
2. Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. In *21st International Conference on World Wide Web*.
3. Xianpei Han, Le Sun, and Jun Zhao. Collective entity linking in web text. In *34th Annual ACM SIGIR Conference*, 2011.
4. Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Furstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Empirical Methods in Natural Language Processing 2011*, pages 782–792, July 2011.
5. David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. In *ACM SIGKDD 2004*, pages 593–598, 2004.
6. Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text. In *ACM SIGKDD*, 2009.
7. Paul McNamee and Hoa Trang Dang. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference*, volume 17, pages 111–113, 2009.
8. Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2, 2014.
9. Bianca Pereira, Nitish Aggarwal, and Paul Buitelaar. AELA: An Adaptive Entity Linking Approach. In *World Wide Web companion*, 2013.
10. Bianca Pereira, Joao C da Silva, and Adriana Vivacqua. Discovering names in linked data datasets. In *1st International Workshop on Web of Linked Entities*, 2012.
11. Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.
12. Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 116–127.
13. Pinar Wennerberg, Klaus Schulz, and Paul Buitelaar. Ontology modularization to improve semantic medical image annotation. *Journal of Biomedical Informatics*, 44(1):155–62, February 2011.