AELA: an Adaptive Entity Linking Approach

Bianca Pereira
DERI, NUIG
Lower Dangan
Galway, Ireland
bianca.pereira@deri.org

Nitish Aggarwal
DERI, NUIG
Lower Dangan
Galway, Ireland
nitish.aggarwal@deri.org

Paul Buitelaar
DERI, NUIG
Lower Dangan
Galway, Ireland
paul.buitelaar@deri.org

ABSTRACT

The number of available Linked Data datasets has been increasing over time. Despite this, their use to recognise entities in unstructured plain text (Entity Linking task) is still limited to a small number of datasets. In this paper we propose a framework adaptable to the structure of generic Linked Data datasets. This adaptability allows a broader use of Linked Data datasets for the Entity Linking task.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

Entity Linking; Linked Data; Named Entity

1. INTRODUCTION

The Entity Linking task is concerned with recognising entity mentions in a text, similar to Named Entity Recognition, but additionally also to link them with respective record(s) in an external database. Since its advent, Wikipedia has played a major role in providing background knowledge for this task [2], but attention has recently shifted towards using Linked Data (LD) instead. Wikipedia contains semi-structured information and requires an effort to extract the links between entities, on the other hand, LD datasets are already structured and allow a more straightforward way to find entities and relationships between them.

Tools such as DBPedia Spotlight [4] and AIDA [6] have been created to benefit as much as possible from LD structure and the amount of data available in the LD cloud. As LD datasets vary in schema, these tools are becoming very specialised for particular datasets. Due to this, only a small number of LD datasets available on the Web is effectively used for the Entity Linking task.

A broader use of available LD datasets would allow us to enable domain-specific identification of entities besides allowing the usage of non-public data such as Linked Enterprise Data. Instead of creating a specialised tool for each available dataset we aim to have a general self-adaptive one that can perform Entity Linking with different LD datasets under varied schemas. For this we created AELA, an Adaptive Entity Linking Approach.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

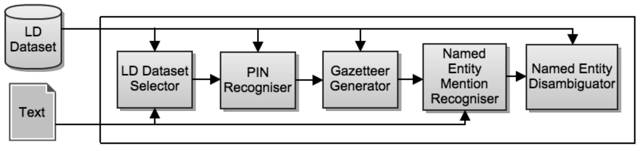

Figure 1: AELA framework

2. OUR APPROACH

AELA is a framework consisting of different modules that perform each step needed to conduct the Entity Linking task. Each module was designed to be independent and adaptable to the structure of the given LD dataset. All modules can be seen in Figure 1.

LD dataset Selector: The first module is responsible for verifying the suitability of the LD dataset for the Entity Linking task. It verifies if both text and dataset share the same domain (music, films and so on) and performs a quality assessment of the dataset.

The assessment is based on three criteria given by [1]: Accessibility, Comprehensibility and Validity of Documents. We also created the Data Richness criterion with two indicators: number of classes with Named Entities as their instances and the number of relationships between entities in the dataset.

PIN Recogniser: The second module adapts the framework with the schema of the selected LD dataset. It detects which classes have Named Entities as their instances and which properties refer to their names. These properties we call PIN (Properties that Identify Names) and each class may have its own set of PIN.

To recognise PIN we are assuming that Named Entity labels are proper names or acronyms with almost all letters capitalised. For this we applied the method and heuristics presented in [5].

Gazetteer Generator: To recognise the Named Entities in a text, a dictionary of Named Entities is required for mapping the names to LD resources. Thus the Gazetteer Generator transforms the LD dataset dynamically into a dictionary through a Lookup Service. This service performs a series of SPARQL queries to find all the resources that refer to a Named Entity with the name provided as input.

Named Entity Mention Recogniser: In order to find each piece of text (potential terms) that mentions a Named Entity, this module uses a sliding window over the text. The dictionary is used to identify names and to link them to a set of candidates in LD resources.

Table 1: Gazetteer Generator results

Dataset	Query Type	#URIs	Correct URIs
Jamendo	simple	94	100%
	RE	93	100%
	full-text	93	100%
Linked MDB	simple	193	100%
	RE	193	100%
	full-text	192	100%

Named Entity Disambiguator: The last module chooses the most appropriate candidate resource to link with each potential term appearing in the text. The Named Entity Disambiguator was developed using a graph-based method adapting the work of [3] to use LD datasets. This decision was based on the fact that some LD datasets do not provide any textual description to allow the applicability of text similarity approaches (e.g. bag-of-words).

3. EVALUATION

The evaluation process assessed each module independently to verify how they perform individually. The Gazetteer Generator received the list with correct PIN as input and the other modules received the output of the previous modules as input.

The framework was applied in two different domains: music and films. For the music domain we collected 66 biographies from the Jamendo website[1] to annotate with the Jamendo RDF Server[2]. For the films domain we selected 28 trivia texts from the Internet Movie Database[3] to use with the Linked Movie Database (Linked MDB)[4].

LD Dataset Selector: Some sources of information were used for quality assessment: the website for each LD dataset, CKAN website[5] (a website for open data datasets) and provided SPARQL endpoints. Both datasets obtained positive results for most indicators so they can be applied for the Entity Linking task. Despite this, some data quality problems such as wrong names, or text instead of an URI as a value for a property, may influence the results of the whole framework.

PIN Recogniser: The method used for this module was exactly the same as in [5] showing that even simple heuristics are capable of identifying PINs.

Gazetteer Generator: We evaluated different kinds of SPARQL queries: simple SPARQL queries, Regular Expression (RE) and full-text search (only available in few SPARQL endpoints). To verify the correctness of the whole Gazetteer requires an extensive manual effort instead we opted for selecting a list of 100 names for each dataset and verify if the Lookup Service returns a set of URIs of Named Entities identified by these names. The results can be seen in Table 1.

Named Entity Mention Recogniser: To evaluate the recognition of Named Entity mentions in text, the result of the module was compared with a manual annotation of both biography and trivia texts. As many people may write texts wrongly and do not write all proper names with capitalised letters, the evaluation was made considering both the whole text and only proper names with capitalised letters. The results are shown in Table 2.

Table 2: Named Entity Mention Recogniser results

Dataset	Type	Names per Text (Avg.)	Correct Names (Avg.)
Jamendo	full	42.32	8.77%
	proper name	8.41	42.57%
Linked MDB	full	7.61	82.52%
	proper name	7.04	87.22%

Named Entity Disambiguator: A manual annotation of texts with the correct URIs from LD resources was performed to evaluate the Named Entity Disambiguator module. Only perfect matching and correct URIs were considered as correct. The results are available in Table 3.

Table 3: Named Entity Disambiguator results

Dataset	Precision	Recall	F-Score
Jamendo	0.42	0.75	0.54
Linked MDB	0.77	0.99	0.87

4. CONCLUSIONS AND FUTURE WORK

In this paper we presented AELA, an adaptive Entity Linking approach using generic Linked Data datasets. Each module from AELA was presented and evaluated independently with texts and LD datasets from two different domains. A following study has to be made to verify how errors propagate from one module to another.

5. ACKNOWLEDGMENTS

This work has been funded in part by the European Union under Grant Number 258191 for the PROMISE project, as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

6. REFERENCES

[1] A. Flemming and O. Hartig. Quality criteria for linked data sources. *Online at http://bit. ly/ld-quality.*

[2] B. Hachey, W. Radford, and J. Curran. Graph-based named entity linking with wikipedia. *Web Information System Engineering–WISE 2011*, pages 213–226, 2011.

[3] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *SIGIR*, 2011.

[4] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *I-Semantics*, 2011.

[5] B. Pereira, J. C. da Silva, and A. S. Vivacqua. Discovering names in linked data datasets. In *First International Workshop on Web of Linked Entities (WoLE 2012)*, 2012.

[6] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12), 2011.

[1] http://www.jamendo.com/

[2] http://dbtune.org/jamendo/

[3] http://www.imdb.com/

[4] http://www.linkedmdb.org/

[5] http://www.ckan.net/