Ontology-Based Information Extraction from Twitter

Kamel Nebhi
LATL, Department of linguistics
University of Geneva, Switzerland
kamel.nebhi@unige.ch

ABSTRACT
The popular microblogging service Twitter provides a vast amount of short messages that contains interesting information for Information Extraction tasks. This paper presents a rule-based system for the recognition and semantic disambiguation of named entities in tweets. As our experimental results shows, performance of this approach measured through BDM looks promising when using Linked Data as Freebase and syntactical context for disambiguation.

KEYWORDS: Ontology-based Information Extraction, Disambiguation, Twitter, Linked Data.

1 Introduction

Twitter is a microblogging service that plays an increasingly important role as a social network and a news media. Twitter allows users to post 140-characters messages and follow users of the world. In February 2012, Twitter had approximately 200 millions active registered users that send 175 millions tweets each day[1]. The growing popularity of Twitter is producing a vast amount of short messages (tweets) that contains interesting information for NLP tasks like Information Extraction (IE).

In this paper, we present an Ontology-based Information Extraction (OBIE) system for Twitter messages using a rule-based approach. Our system combines Named Entity Recognition (NER) and a disambiguation module. We show that our system can improve disambiguation efficiency using syntactical context and knowledge base like Freebase.

The paper is divided as follows. First we present our approach in Section 2. Then, we present the experimental setup for testing in Section 3. Finally, we summarize the paper.

2 Approach

2.1 OBIE

Information Extraction is a key NLP technology to introduce complementary information and knowledge into a document. The term "Ontology-based Information Extraction" has been conceived only a few years ago and has recently emerged as a subfield of IE. OBIE is different from traditional IE because it finds type of extracted entity by linking it to its semantic description in the formal ontology. The task of OBIE has received a specific attention in the last few years with many publications that describe systems.

For (Wimalasuriya and Dou, 2010), the key characteristics of OBIE systems are:

- *Process natural language text documents*: the inputs of OBIE systems are limited to unstructured or semi-structured documents.

[1] http://www.creativethinkingproject.com/wp-content/uploads/2012/03/CTP_SocialMediaStatistic.pdf

Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 17–22,
COLING 2012, Mumbai, December 2012.

- *Information Extraction process guided by an ontology*: the information extraction process is guided by the ontology to extract things such as classes, properties and instances.
- *Present the output using ontologies*: OBIE systems must use an ontology to represent the output.

2.2 Architecture

Our OBIE system is built on GATE (Cunningham et al., 2011) to annotate entities in tweets and relate them to the DBpedia ontology[2] where appropriate. The DBpedia ontology is a shallow, cross-domain ontology, which has been manually created based on the Wikipedia projects. The ontology organizes the knowledge according to a hierarchy of 320 classes and 1650 different properties.

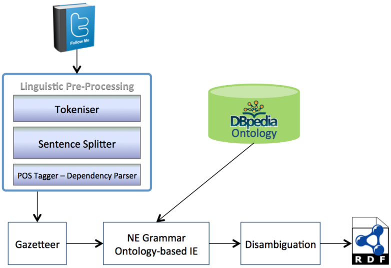

Figure 1: Ontology-based Information Extraction Architecture

Figure 1 describes the architecture of our OBIE system. The source data is a set of short messages from *BBC News*, *New York Times* and *The Times* Twitter accounts. Semantic annotation is performed by GATE with respect to the DBpedia ontology. The GATE application consists of a set of processing resources executed in a pipeline over a corpus of documents. The pipeline consists of 5 parts:

- Linguistic pre-processing
- Gazetteer (used to identify entities directly via look-up)
- Rule-based semantic annotation
- Disambiguation
- Final output creation

The linguistic pre-processing phase contains GATE components such as tokenisation and sentence splitter. It also contains specific tools like Stanford Parser for English part-of-speech tagging and dependency parsing. The gazetteer lookup phase comprises combination of default gazetteer lists from ANNIE[3] and some newly gazetteer lists extract from Wikipedia and DBpedia. The grammar rules for creating semantic annotation are written in a language called JAPE, which is a finite state transducer. The rules are based on pattern-matching using several information taken from the gazetteer and the Stanford Parser results.

[2] `http://wiki.dbpedia.org/Ontology`

[3] GATE is distributed with an IE system called ANNIE (A Nearly-New IE system). It comprises a set of core processing like tokeniser, sentence splitter, POS tagger, Gazetteers, JAPE transducer, etc.

2.3 Disambiguation process

Disambiguation process consists in determining which sense of a word is used when it appears in a particular context. For example, the string "Washington" is used to refer to more than 90 different named entities in DBpedia database. Our approach use popularity score and syntax-based similarity to disambiguate each surface form (Hoffart et al., 2011).

Popularity Score: This score is provided by Freebase API (Bollacker et al., 2008) [4]. Popularity score of entities can be seen as a probabilistic estimation based on Wikipedia frequencies in link anchor. The search Freebase API allows access to Freebase data given a text query. A large number of filter constraints are supported to better aim the search at the entities being looked for.

For example, this GET request `https://www.googleapis.com/freebase/v1/search?query=Washington&filter=(any%20type:/people/person)&indent=true` only matched persons named "Washington".

Syntax-based Similarity: After selecting candidate entities for each surface form with Freebase, our system uses the context around the surface forms. For this, we construct a context from all words in the short messages using syntactic information based on dependency trees (Thater et al., 2010). We created rules to determine the immediate syntactic context in which an entity mention occurs.

Table 1 shows examples of NER for a BBC News tweet. For this short message, the system use popularity score to disambiguate entities.

Tweet	@BBCBreaking: Hamas says senior military figure Ahmed al-Jabari killed in Israeli airstrike in Gaza.
Extracted Mentions	Hamas, Ahmed Jabari, Gaza
Candidate Entities	{Hamas, Hamas of Iraq}, {Ahmed Jabari, Abdullah Muhammed Abdel Aziz}, {Gaza, Gaza strip, Palestinian territories}

Table 1: Example of Extraction for a BBC News Tweet

3 Evaluation

3.1 The "Balanced Distance Metric"

Traditional IE systems are evaluated using Precision, Recall and F-Measure. These measures are inadequate when dealing with ontologies. In (Maynard et al., 2008), Maynard analyzed several metrics and proved that the Balanced Distance Metric (BDM) is useful to measure performance of OBIE systems taking into account ontological similarity. In fact, BDM is a real number between 0 and 1 depending on several features like:

- The length of the shortest path connecting the two concepts
- The depth of the two concepts in the ontology
- Size and density of the ontology

[4] Freebase is a collaborative knowledge base used to structure general human knowledge. It contains structured data of almost 23 million entities.

The BDM itself is not sufficient to evaluate our system, so we decide to combine the traditional Precision and Recall with the BDM measure to obtain an Augmented Precision (AP) and an Augmented Recall (AR) as follows :

$$AP = \frac{BDM}{n + Spurious^{5}} \quad and \quad AR = \frac{BDM}{n + Missing^{6}}$$

In order to demonstrate the usefulness of this measure, Table 2 presents some examples of entities misclassified by the system. All the concepts involved in Table 2 are illustrated in Figure 2, which presents a part of the DBpedia ontology.

In the first example, the predicted label PopulatedPlace is 2 concepts away from the key label City but its BDM_F_1 value is lower than the corresponding value in the fourth example, mainly because the concept PopulatedPlace occupies a higher position in the DBpedia ontology than the concept BodyOfWater. BDM also considers the concept densities around the key concept and the response concept. We can show the difference by comparing the second and the third example. They have the same predicted label Company, and their key labels MediaCompany and RecordLabel are two sub-concepts of Organisation. However, their BDM values are different. This is because the concept MediaCompany has one child node but the concept RecordLabel has no child node.

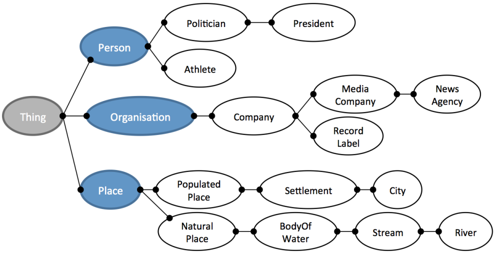

Figure 2: View of the DBpedia Ontology

Entity	Result	Key	BDM_F_1
Boston	PopPlace	City	0.43
Vivendi	Company	MediaCompany	0.55
UMG	Company	RecordLabel	0.45
Ourcq	BodyOfWater	River	0.45

Table 2: Examples of entities misclassified by the system

[5] The application marked an entity that is not annotated in the key (=False Positive).

[6] The entity is annotated in the key but the application failed to mark it (=False Negative).

3.2 Experiment

To evaluate the performance of the system we applied the processing resources on the evaluation corpora of 115 short messages from *BBC News*, *New York Times* and *The Times* Twitter accounts. We manually annotated these documents with the concepts of the DBpedia ontology. Then, we compare the system with the gold standard.

For the evaluation, we only use Person, Organization and Location named entity categories. In table 3, the system without disambiguation achieved a traditional F-Measure of 65% and an augmented F-Measure of 69%. Adding the disambiguation layer improve extraction effectiveness, traditional F-Measure rises to 86% and augmented F-Measure rises to 90%.

	F_1	BDM_F_1
OBIE (no disambiguation)	0.65	0.69
OBIE (with disambiguation)	0.86	0.90

Table 3: Results

Conclusion and perspectives

In this paper we introduced an approach for Ontology-based Information Extraction from Twitter. Our system provides an integrated disambiguation module based on popularity score and syntax-based similarity. As our evaluation shows, the system performed significantly better using disambiguation process.

In future work, we plan to incorporate more knowledge from Linked Data and we'll integrate the application into a ReSTful Web service.

References

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, USA. ACM.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792.

Maynard, D., Peters, W., and Li, Y. (2008). Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In *LREC*.

Thater, S., Fürstenau, H., and Pinkal, M. (2010). Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 948–957, Stroudsburg, PA, USA.

Wimalasuriya, D. C. and Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *J. Information Science*, 36(3):306–323.