EXTRACTION OF SEMANTIC RELATIONS FROM TEXT

Petr Knoth
Master Degree Programme (1), FIT BUT
E-mail: xknoth00@stud.fit.vutbr.cz

Supervised by: Pavel Smrž
E-mail: smrz@fit.vutbr.cz

ABSTRACT

In recent years the amount of unstructured data stored on the Internet and other digital sources has increased significantly. These data contain often valuable, but hardly retrievable information. The term *unstructured data* refers mainly to data that does not have a data structure. As a result of this, the unstructured data is not easily readable by machines. In this work, we present a simple method for automatic extraction of semantic relations that can be used to precisely locate valuable pieces of information.

1 INTRODUCTION

Information Extraction (IE) is usually defined as the process of selectively structuring and combining data that are explicitly stated or implied in one or more documents. This process involves a semantic classification of certain pieces of information and is considered as a light form of text understanding [2]. The structured information can be then in turn used as a basis for question answering, machine translation, semantic web systems etc. Currently, there is a considerable interest in using these systems for information retrieval. This is caused by an increasing need to localize precise information, rather than just retrieving a list of the most relevant documents.

2 PATTERN RECOGNITION

In this work we are concerned with written natural language texts. Our assumption is that natural language texts are not completely irregular and that is why we are able to identify common patterns, which can serve as a first step for retrieving the semantics of a language. As it was learned in the foregoing parts, information extraction relies on pattern recognition methods. Pattern recognition (also known as classification or pattern classification) aims at classifying data (patterns) based on either a priori knowledge that is acquired by human experts or on knowledge automatically learned from data. A system, that automatically sorts patterns into classes or categories is called a *pattern classifier* [2]. The classification patterns consists of features and their values. In our case, the features are textual characteristics that can be identified or measured, and that are assumed to have discriminative value.

In the next section we describe an experiment that was performed on a selection of Wikipedia articles. Our goal was to precisely identify pairs *country - capital city* in an unstructured text. This task looks similar to one defined in [1], where the goal is to extract relation pairs, such as *organization - location* from plain text. In [1] the proposed Snowball system is expected to extract these relations between two entities that are both identified in a text. However, we are here interested in finding relations between the whole article, which refers to a certain country,

and possibly a few capital city entities. First of all, we will define the task and the *Ideal* set, which is a set of pairs *country - capital city* we want to retrieve. Then we briefly explain the method and evaluate the system using standard precision and recall metrics.

3 A PROTOTYPE SYSTEM FOR INFORMATION EXTRACTION

3.1 DATA SET

A new data set was created from a selection of Wikipedia articles in order to evaluate the performance of information extraction algorithms. The data set contains about 1.6 MB of text stored in 50 files. Each file corresponds to a particular country. Only articles, which at least once mention the name of their national capital city, were selected. By the word "mention" we refer to all parts of the article from which a human can recognize the name of a capital city.
All the articles were tokenized and sentences were splitted. Furthermore, all files were manually inspected and sentences, which can serve as a clue to determine the name of a capital city, were annotated. Finally, these capital city entities together with the names of their corresponding countries form a set of positive examples. We will denote this set as an *Ideal* set. Although the whole data set contains only 99 of these positive example pairs, it also contains a range of sentences that seem quite problematic. We discuss them in section 3.2.

3.2 METHOD

The system is initially provided with a number of example pairs country - capital. The expected output are new pairs country - capital extracted from the data set. We assume we can decide if a word is a capital city for a given country by exploring its context. The evaluated approach uses only lexical features weighted according to their importance for the task.
The **first step** is to search for all occurrences of a capital, provided by an example from the training set, in a corresponding text. Its context words are then explored. Note that the size of a context window is fixed during an experiment.
We distinguish context words appearing on the left hand side of the entity from context words appearing on the right hand side. Context words are also weighted using a simple idea similar to *term frequency-inverse document frequency (tfidf)*: most indicative words tend to appear often in a context of capital city, but they rarely appear in different contexts. Using this approach it was for example quite easy to automatically detect that the most indicative word for this task is the word "capital." Summarizing the paragraph, weighted words on the left and right hand side of a capital city entity form a pattern. We refer to those patterns as candidate patterns.
The **second step** is to prune the set of candidate patterns. Our candidate patterns can contain patterns extracted from sentences that mention the name of a capital city *A* (*A* is a member of the training set), but it is impossible from them to conclude that *A* is the capital city. For example, it is incorrect to derive that Prague is capital city from the sentence: *The occupation ended on 9th May 1945 with the arrival of Soviet and American armies and the **Prague** uprising.* Patterns extracted from such sentences can cause errors while identifying new entities. Therefore, minimal confidence threshold is used to discard them. Candidates with low confidence are considered unreliable and are eliminated from further evaluation. Other candidates are considered reliable and form a pattern set.
The **third step** is to use the learned patterns in the pattern set for identification of new relations from unseen articles. The algorithm scans the text of an article word by word. All words beginning with a capital letter are considered as potential candidates for new capital city entities. Their left hand side context words together with the right hand side context words form a pattern

as in the first step. This pattern is compared with all patterns in the pattern set. A pattern similarity threshold controls how flexible the patterns are in identifying new entities.

4 RESULTS

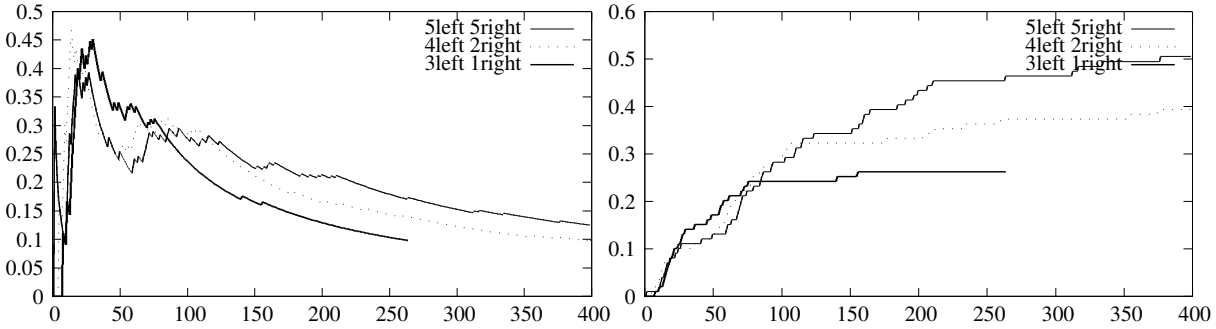

Figure 1: (a) Precision (y-axis) / discovered pairs (x-axis), (b) Recall / discovered pairs

Given the *Ideal* set of sentences from which a relation can be derived we define *precision* and recall as

$$precision = \frac{\sum_{i=0}^{|Extracted|} |l_i = l_i'|}{|Extracted|} \quad Recall = \frac{\sum_{i=0}^{|Extracted|} |l_i = l_i'|}{|Ideal|} \tag{1}$$

where $Extracted$ is a set of relation pairs that were extracted by the system and $[l_i = l_i']$ is equal to 1 if the test value l_i matches the extracted value l_i' and equal to 0 otherwise.
Thus, *precision* refers to a proportion of correctly identified relation pairs to the size of the extracted pairs and *recall* to a proportion of correctly identified pairs to the whole *Ideal* set.
The *leave-one-out cross-validated (LOOCV)* results are reported in Figure 1. The algorithm performed relatively well on the first one hundred extracted items. Then the precision is decreasing and it is evident that it is quite hard for the method to identify new entities. This may be caused by a limited size of the context window that is maybe still too small to cover all important context words. The results show, that we were able to discover more patterns using a bigger size of the context window.
Here we present an example sentence from which the relation was correctly derived: *The Government of Kazakhstan transferred its capital from Almaty to* ***Astana*** *on December 10 1997* On the contrary, typical mistakes were derived from sentences such as: ***Homel*** *with 481000 people is the second largest city of Belarus and serves as the capital of the* ***Homel*** *Oblast.*

5 CONCLUSION

We have developed a prototype system for pattern extraction and evaluated its performance on the selection of Wikipedia articles. This system can now serve as a baseline and as an evaluation framework for comparison with advanced machine learning techniques that are about to be developed in the following months.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[2] Marie-Francine Moens. *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Springer, October 2006.