

Information Extraction tasks: a survey

Gonçalo Simões, Helena Galhardas, Luísa Coheur

Instituto Superior Técnico, INESC-ID, DMIR, L2F

Abstract. An Information Extraction activity is a complex process that can be decomposed into several tasks. This decomposition brings the following advantages: *(i)* for each task it becomes possible to choose the best technique independently from the other tasks; *(ii)* an Information Extraction program can be developed as a set of independent modules (one for each task), making it easy to perform local debugging; *(iii)* it becomes easy to customize the Information Extraction activity through reordering, selection or even composition the tasks.

This paper presents a commonly used decomposition of the Information Extraction activities and gives detail about the most used machine learning and rule-based techniques for each task.

Keywords: Information Extraction, Natural Language Processing, Machine Learning

1 Introduction

With the increasing volume of publicly available information, companies need to develop processes for mining information that may be vital for their business. Unfortunately, much of this information is presented in the form of unstructured or semi-structured texts. Software tools are not able to analyze such texts and humans would take so much time to perform this task that the information would become obsolete by the time it was available.

Information Extraction emerged as a solution to deal with this problem. According to (Cowie & Lehnert, 1996), “*Information Extraction* starts with a collection of texts, then transforms them into information that is more readily digested and analyzed. It isolates relevant text fragments, extracts relevant information from the fragments, and then pieces together the targeted information in a coherent framework”.

An *Information Extraction* activity can be very complex. Thus, it is common to decompose it into several tasks. This decomposition offers some advantages. First, it is possible to choose, for each task, the techniques and algorithms that better fit the objective of a particular application. Second, it is easy to locally debug an *Information Extraction* program since the module responsible for each task is completely independent from the others. Finally, an *Information Extraction* can be customized activity according to an application’s needs, by reordering, selecting and composing some of the tasks.

This paper presents a possible decomposition of the *Information Extraction* activity in tasks. This decomposition is based on the work of (McCallum, 2005).

The considered tasks are: Segmentation, Classification, Association, Normalization and Correference Resolution. For each of these tasks we explain what is its purpose and which techniques can be used. An extended version of this paper can be found in (Simões, Galhardas, & Coheur, 2009).

2 Segmentation

The *Segmentation* task divides the text into atomic elements, called segments or tokens. Even though this task is simplified for Western languages due to the existence of whitespaces separating words, there are some cases in which simple whitespace separation may not be enough (Santos, 2002). Usually, segmentation for these cases is performed using rules that show how to handle each case.

The major problems related to this task can be found in oriental languages. For example, the Chinese does not have whitespaces between words (Li Haizhou & Zhiwei, 1998). For this reason, solving the problems described above is not enough in this language. In these cases, it is typically necessary to use external resources. Lexicons and grammars can also be used in order to accomplish the task of segmentation using syntactic or lexical analysis. Another approach for segmentation in Chinese uses techniques based on statistics. An example is the system described in (Li Haizhou & Zhiwei, 1998), which uses *N-grams* and the *Viterbi algorithm* (Forney, 1973) applied to segmentation.

3 Classification

The *Classification* task determines the type of each segment obtained in the segmentation task. In other words, it determines the field of the output data structure where the input segment fits. The result of this task is the classification of a set of segments as entities, which are elements of a given class potentially relevant for the extraction domain.

The rule-based techniques used in the classification task are usually based on linguistic resources, such as lexicons and grammars (Farmakiotou et al., 2000).

One of the most popular approaches to undertake classification is machine learning. Machine learning techniques used in this task are usually supervised, which means that an annotated corpus is needed. Five of the most common supervised learning techniques are the Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM) (McCallum, Freitag, & Pereira, 2000), Conditional Random Fields (CRF) (Lafferty, McCallum, & Pereira, 2001), Support Vector Machines (Isozaki & Kazawa, 2002) and Decision Trees (Sekine, Grishman, & Shinnou, 1998).

4 Association

The *association* task seeks to find how the different entities found in the classification task are related. The systems that perform extraction of relationships are less common than the ones that perform the classification task (McCallum, 2005). This happens due to the difficulty in achieving good results in this task.

Many techniques in the association task are based on rules. The simplest approach uses patterns to extract a limited set of relationships. A more generic

rule-based approach for association is based on *syntactic analysis*. Often, the relationships that we want to extract are grammatical relationships (Grishman, 1997). For example, a verb may indicate a relationship between two entities.

The association task can also use machine learning techniques. One of the first machine learning approaches was based on probabilistic context-free grammars (Miller et al., 1998). These grammars differ from regular context-free grammars, because they have a probability value associated to each rule. When the syntactic analysis is undertaken, it is possible to find many syntactic trees. By using probabilistic rules, the probability of each tree is computed and the most probable tree is chosen.

5 Normalization and Coreference Resolution

Normalization and Coreference resolution are the less generic tasks of the *Information Extraction* process (Applet & Israel, 1999), since they use heuristics and rules that are specific to the data domain.

The *normalization* task is required because some information types do not conform to a standard format. This task is typically achieved through the use of conversion rules that produce a standard format previously chosen.

Coreference arises whenever the same real world entity is referred in different ways in a text fragment. This problem may arise due to the use of: *(i)* different names describing the same entity (e.g., the entity “Bill Gates” can be found in the text as “William Gates”), *(ii)* classification expressions (e.g., a few years ago, “Bill Gates” was referred as “the world’s richest man”), *(iii)* pronouns (e.g., in the sequence of sentences “Bill Gates is the world’s richest man. He was a founder of Microsoft”, the pronoun “He” refers to “Bill Gates”).

Rule-based approaches for coreference usually take into account semantic information about entities. A machine learning approach for coreference resolution is described in (Cardie & Wagstaff, 1999). This approach is based on clustering algorithms for grouping similar entities.

6 Conclusions

In this paper, we presented the decomposition of an *Information Extraction* activity into tasks and referred the techniques that are commonly used for each task. The goal of this paper is to supply material that can be used for the conceptualization of an *Information Extraction* program.

References

- Applet, D., & Israel, D. (1999). Introduction to information extraction technology. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden.
- Cardie, C., & Wagstaff, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint Sigdat Conference on empirical methods in natural language processing and very large corpora* (pp. 82–89). New Brunswick, NJ, USA.

- Cowie, J., & Lehnert, W. (1996). Information extraction. In *Special natural language processing issue of the communications of the ACM* (Vol. 39, pp. 80–91). New York, NY, USA.
- Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)* (pp. 75–78). Pyrgos, Greece.
- Forney, G. D. (1973). The Viterbi algorithm. In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)* (Vol. 61, pp. 268–278).
- Grishman, R. (1997). Information extraction: techniques and challenges. In *Information Extraction International Summer School SCIE-97* (pp. 10–27). Frascati, Italy.
- Isozaki, H., & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING02)* (pp. 390–396). Taipei, Taiwan.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)* (pp. 282–289). Williamstown, MA, USA.
- Li Haizhou, B. S., & Zhiwei, L. (1998). Chinese sentence tokenization using viterbi decoder. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP 1998)*. Singapore.
- McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. In *ACM Queue* (Vol. 3, pp. 48–57). New York, NY, USA.
- McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)* (pp. 591–598). Stanford, CA, USA.
- Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., et al. (1998). Algorithms that learn to extract information—BBN: Description of the SIFT system as used for MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7)* (pp. 75–89). San Francisco, CA, USA.
- Santos, M. (2002). *Extraindo regras de associação a partir de textos*. Mestrado em Informática Aplicada, Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.
- Sekine, S., Grishman, R., & Shinnou, H. (1998). A decision tree method for finding and classifying names in japanese texts. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)* (pp. 171–178). Montreal, Canada.
- Simões, G., Galhardas, H., & Coheur, L. (2009). *Information extraction tasks: a survey* (INESC-ID technical report No. 37/2009). Lisbon, Portugal.