Lexical Knowledge Acquisition Using Spontaneous Descriptions in Texts

Augusta Mela, Mathieu Roche, Mohamed El Amine Bekhtaoui

Lexical Knowledge Acquisition Using Spontaneous Descriptions in Texts

Augusta Mela[1], Mathieu Roche[2], and Mohamed El Amine Bekhtaoui[3]

[1] Univ. Montpellier 3, France
[2] LIRMM – CNRS, Univ. Montpellier 2, France
[3] Univ. Montpellier 2, France

Abstract. This paper focuses on the extraction of lexical knowledge from text by exploiting the "glosses" of words, i.e. spontaneous descriptions identifiable by lexical markers and specific morpho-syntactic patterns. This information offers interesting knowledge in order to build dictionaries. In this study based on the RESENS project, we compare two methods to extract this linguistic information using local grammars and/or web-mining approaches. Experiments have been conducted on real data in French.

1 Introduction

Automatic acquisition of lexical knowledge from texts aims to detect various types of lexical units (e.g. terms, named entities, phrases, new words, and words with a new meaning) as well as their syntactic and semantic properties. In a multilingual context, using bilingual texts, automatic acquisition consists in detecting the translations of these units. Automatic acquisition constitutes a precious help in order to build dictionaries, thesaurus, and terminologies for general and/or specialized domains [1]. This task is also useful for documentary research thanks to query expansion.

Our work concerns lexical structuration. To situate it, we will restrict ourselves to work on lexical structuration. We adopt the distinction proposed by [2] concerning the different levels of structuration:

1. Semantic lexical relationships such as synonymy or hypernymy concern microstructuration,
2. The classification of units in topics concerns macrostructuration: At this level the nature of the link between units is not identified.

In macrostructuration, the existing approaches are mainly based on Harris distributional semantics [3]. According to this theory, the set of the contexts in which the word appears allows to draw its portrait and to determine the meaning. In practice, given a word, we look for which words we can substitute for it (paradigmatic properties), with which words it can be combined (syntagmatic properties), or more loosely in what immediate context it appears (co-occurrence properties). These questions are not new and lexicographers have always used

G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 366–371, 2012.

corpora to answer them. What has changed today is the size of the corpora and the computer tools capable of systematizing the distributional approach.

In microstructuration, the approaches are various:

- Some approaches use the structure of the unit itself (e.g., *le coussin de sécurité* (security cushion) is a hypernym of *coussin de sécurité arrière* (rear security cushion) and *anticapitaliste* is an antonym of *capitaliste*).
- Following the distributional approach, other methods consider that the related units appear in similar context. This one can be limited to the words in the close environment of the target word or can be extended to refer to the entire text. Represented in a word-based vector space, both words are similar if their vector representations are close [4].
- Another type of approach which is close to our work, starts from the principle that semantic relationships are expressed by linguistic markers (lexical or grammatical elements), or paralinguistic elements (punctuation marks, inverted commas). In this approach, morpho-syntactic patterns which combine these marks are manually built and projected into the texts to seek out the items supposedly in the semantic relationships under study. These patterns can also be acquired semi-automatically [5]. The quality of the results of this approach depends on the precision with the patterns have been defined or the quality of the corpus used to infer them.

Section 2 details our approach which combines linguistic and statistic information. Section 3 presents the obtained results, and Section 4 outlines future work.

2 The Word and Its Gloss

2.1 Definition

In the RESENS project (PEPS-CNRS project), linguists and computer scientists work together to study the phenomena of the gloss and to propose methods to access the meanings of words. Glosses are commentaries in a parenthetic situation, often introduced by markers such as *appelé*, *c'est-à-dire*, *ou* (called, that is to say, or) which signal the lexical semantic relationship involved: Equivalence, with *c'est-à-dire*, *ou* ; specification of the meaning, with *au sens* (in the sense); nomination, with *dit*, *appelé* (called); hyponomy, with *en particulier*, *comme* (in particular, like); hypernymy, with *et/ou autre(s)* (and/or other), etc. They are in apposition to the glossed word, usually of the nominal category. Their syntactic and semantic topologies have been established by linguists [6]. It has also been noted that their frequency depends on the genre of the texts: They are more often found in didactic texts or work of popularization than in poetry.

A gloss is spontaneous. It shares this characteristic with a so-called natural definition because neither is the fruit of the reflective work of a lexicographer. However, a gloss is parenthetical whereas a natural definition is the main object of the proposition. Thus the gloss can roughly be described by the configuration X marker Y as described in the following section.

2.2 Gloss Extraction

In this section we detail our method, considering the case of glosses with *appelé*. We start from the principle that X and Y are noun phrases (NP) extracted with the approach of [7]. In addition, we take into account that Y can be a coordination of NPs, thus the variant of the abstract pattern of the gloss becomes: X marker $Y_1, Y_2...Y_n$.

- **Local Pattern 1:** The first pattern detects Y_1, the first NP to the right of the marker. For example, the sentence *Un disque microsillon, appelé disque vinyle* (a grammaphone record, called vinyl record) allows to extract X = *un disque microsillon* and Y_1 = *disque vinyle*.

- **Local Pattern 2:** A second pattern takes into account the possibility of coordination in position Y, to extract a sequence such as *un disque microsillon appelé disque vinyle ou Maxi*. Two NPs are extracted: Y_1 = *disque vinyle* and Y_2 = *Maxi*.

- **Global Extraction:** Once the gloss with *appelé* had been detected in a corpus, we look for the NPs situated between the marker and a right hand boundary. This right boundary is either a conjugated verb or a strong punctuation mark.

The following section describes a ranking function which orders the extracted phrases Y_i.

2.3 Gloss Ranking

Our web mining approach is based on the Dice measure. This one calculates the "dependency" between both Nominal Phrases (NP), i.e. dependency in terms of co-occurrence more or less close. This dependency is calculated by querying the Web.

Applied to X (i.e. glossed NP) and Y_i (glossing NP), the measure is defined with the formula (1):

$$Dice1(X, Y_i) = \frac{2|X \cap Y_i|}{|X|+|Y_i|} \quad (1)$$

where $|X \cap Y_i|$ is the number of web pages containing the words X and Y_i one beside the other, $|X|$ is the number of pages containing the term X, and $|Y_i|$ is the number of pages containing Y_i. In order to calculate $|X \cap Y_i|$ by querying the Web, we use quotations (").

Subsequently, we release the proximity constraints between phrases. So the number of web pages where X and Y_i are present together is calculated. In this

case, we use a measure called $Dice2$. The numerator of this measure (formula (2)), represents the number of times where X and Y_i are in the same pages.

$$Dice2(X, Y_i) = \frac{2|X \ AND \ Y_i|}{|X|+|Y_i|} \quad (2)$$

From both measurements (i.e. $Dice1$, $Dice2$), two types of combinations are proposed.

1. The first one, $Diexact$, is $Dice1$ if $Dice1$ returns a result. Otherwise $Dice2$ is calculated. So this approach gives priority to $Dice1$.

2. The second one, $Dibary$ (formula (3)), calculates the barycenter of $Dice1$ and $Dice2$.

$$Dibary_k(X, Y_i) = k.Dice1(X, Y_i) + (1 - k).Dice2(X, Y_i) \quad (3)$$
$$\text{where } k \in [0, 1]$$

Our web mining methods can use different search engines (e.g. Google, Yahoo, and Exalead) to calculate Dice measures. Experiments based on 22 texts containing glosses have shown that the API provided by Yahoo have a good behavior. We chose this search engine in our experiments.

3 Experiments

Our corpus is composed of 219 candidate NPs (i.e Y_i) identifyied with the word *"appelé"* (i.e. "called"). These candidates are extracted with a software implemented in Java (see Figure 1).

A linguist expert assignes the mark 1 for the relevant pairs (X, Y), and the mark 2 for the very relevant pairs. When the NP is partially extracted, the couple (X, Y) is evaluated with a score at 1. Irrelevant NP are evaluated at 0.

Evaluation of Extraction Methods

The quality of element X are: 6 NP evaluated as relevant (mark 1), 68 NP evaluated as very relevant (mark 2). Now we can compare the methods in order to extract the 219 glossing NP (i.e. Y_i).

Table 1 presents the results obtained (evaluations at 0, 1, or 2). The results show that the majority of phrases are evaluated as very relevant. The global extraction of Y_i produces a quarter of irrelevant phrases. We can explain this situation because the NP following *"appelé"* are not always semantically related to X. Nevertheless, the global extraction can identify a quantity larger of relevant NPs than the use of simple patterns.

With the marks 1 and 2 considered as relevant, precision, recall, and F-measure are presented in Table 1. This table shows that the best F-measure is obtained with the global extraction of phrases. Note that the use of coordination rules (i.e. Local Pattern 2) is effective (excellent precision and good recall).

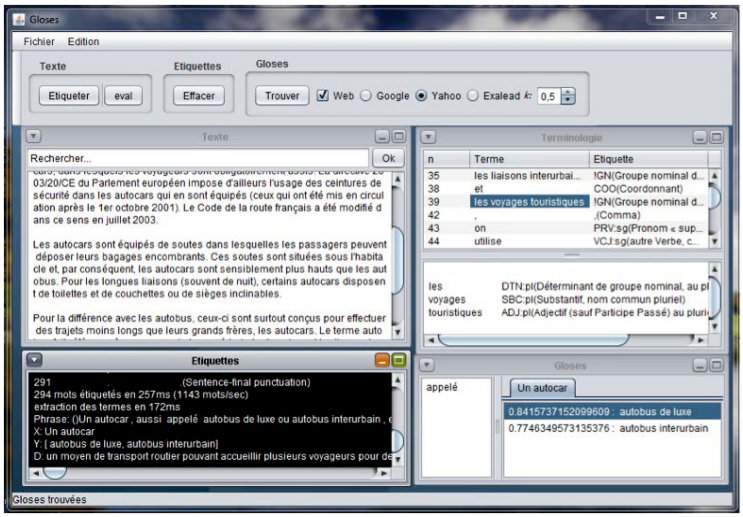

Fig. 1. Extraction of glosses – Software

Table 1. Evaluation of the different methods

Method / Evaluation	Nb of extracted Y_i		
	Evaluation at 1	Evaluation at 2	Evaluation at 0
Local Pattern 1	3 (3.75 %)	**74 (92.50 %)**	3 (3.75 %)
Local Pattern 2	4 (3.67 %)	**101 (92.66 %)**	4 (3.67 %)
Global Extraction	16 (7.30 %)	**150 (68.49 %)**	53 (24.21 %)

Method / Evaluation	Precision	Recall	F-measure
Local Pattern 1	**0.96**	0.46	0.62
Local Pattern 2	**0.96**	0.63	0.76
Global Extraction	0.75	**1**	**0.86**

Evaluation of Ranking Functions

Now we can evaluate the ranking quality of NP obtained by web-mining approaches. To compare our algorithms, we calculate the sum of the ranks of Y_i evaluated as relevant by the expert. Actually the minimization of this sum is equivalent to maximize the Area Under the ROC Curve [8]. This principle is often used in data-mining field to assess the quality of ranking functions. The method giving the best results returns the lower value.

The sum of the ranks of relevant Y_i obtained with our evaluation corpus is shown in Table 2. This one shows that *Diexact* method gives best results. The influence of k for *Dibary* is low.

Table 2. Sum of relevant elements

Method	*Diexact*	*Dibary*					
		$k = 0$	$k = 0.2$	$k = 0.4$	$k = 0.6$	$k = 0.8$	$k = 1$
Sum	**323**	329	329	329	329	329	364

4 Conclusion and Future Work

In this paper, we have presented a method to extract NPs in relationship by a gloss phenomenon. Our methods combine local grammars and statistical associations of units on the web. In the context of the RESENS project, these methods have been manually evaluated on real data.

In our future work, we would like to perform a contrastive analysis of English/French corpora in order to give a new point of view of the phenomenon of spontaneous descriptions. A first study on aligned English/French texts reveals frequent regularities of glosses in a multilingual context. The alignment enables to improve the multilingual lexical acquisition of new words and their translations.

Moreover we plan to test other web mining measures, these ones will be able to take into account other kinds of operators for querying the web (e.g. *Near* operator).

Finally we plan to focus our work on the study of the markers between NPs in order to automatically extract the type of relationships (synonymy, hyponomy, hypernymy, and so forth).

Acknowledgment. We thank Vivienne Mela who improved the readability of this paper.

References

1. Muresan, S., Klavans, J.: A method for automatically building and evaluating dictionary resources. In: Proceedings of LREC (2002)
2. Nazarenko, A., Hamon, T.: Structuration de terminologie: quels outils pour quelles pratiques? TAL 43-1, 7–18 (2002)
3. Daladier, A.: Les grammaires de harris et leurs questions. Languages (99) (1990)
4. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. In: Proc. of ACM, vol. 18, pp. 613–620 (1975)
5. Aussenac-Gilles, N., Jacques, M.P.: Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMELEON. Terminology, Pattern-Based Approaches to Semantic Relations 14(1), 45–73 (2008)
6. Steuckardt, A.: Du discours au lexique: la glose. Séminaire ATILF (2006)
7. Daille, B.: Study and implementation of combined techniques for automatic extraction of terminology. In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pp. 49–66 (1996)
8. Ferri, C., Flach, P., Hernandez-Orallo, J.: Learning decision trees using the area under the ROC curve. In: Proceedings of ICML, pp. 139–146 (2002)