Towards a methodology for automatic identification of hypernyms in the definitions of large-scale dictionary

Inga Gheorghita[1,2], Jean-Marie Pierrel[1]
[1] ATILF, Université de Lorraine & CNRS, Nancy, France
[2] XILOPIX, 2 rue de Nancy, 88000 Épinal, France
inga.gheorghita@atilf.fr, jean-marie.pierrel@atilf.fr

Abstract

The purpose of this paper is to identify automatically hypernyms for dictionary entries by exploring their definitions. In order to do this, we propose a weighting methodology that lets us assign to each lexeme a weight in a definition. This fact allows us to predict that lexemes with the highest weight are the closest hypernyms of the defined lexeme in the dictionary. The extracted semantic relation "is-a" is used for the automatic construction of a thesaurus for image indexing and retrieval. We conclude the paper by showing some experimental results to validate our method and by presenting our methodology of automatic thesaurus construction.

Keywords: hypernymy, dictionary, thesaurus

1. Introduction

Linguistic resources such as dictionaries, computational lexicons, semantic taxonomies and thesauri are an important source of knowledge for natural language processing applications.

The information contained in linguistic resources, depending on their type, includes semantic relations (eg. *thrush* is a kind of *bird*), text definitions (eg. the Oxford dictionary defines the "lion" as "a large tawny-colored cat that lives in prides"), examples on the usage domain, and so on. Unfortunately, not all of which provide structured information that can be used by applications of natural language processing (Harabagiu, Miller, & Moldovan, 1999). A human understands the meaning of a word just by reading its definition in the dictionary, but it's not the case for a computer system. The main cause is that the semantic information, such as definitions, contained in the lexical resources is not very explicit and is provided in the form of free text.

Even WordNet (Miller, 1995), one of the most popular lexicons for the English language, uses definitions to explain the meaning of ambiguous words. However, compared with other electronic dictionaries and thesauri, which represent only an electronic transcription of their paper version, WordNet contains explicit information in the form of semantic relations such as, meronymy and hypernymy.

Over the last several decades much research has been done on the automatic construction of resources from corpora (Hearst, 1992), (Yarowsky, 1992), in particular by creating hypernym hierarchies. Various techniques as Machine Learning (Snow, Jurafsky, & Ng, 2005), Hidden Markov Model (Ritter & Soderland, 2009) and resources as dictionaries (Nakamura & Nagao, 1988) and thesauri (Kennedy & Szpakowicz, 2007) are used for identification of hypernymy or other semantic relations in the text.

In this paper, we aim to make explicit the information that is implicitly contained in the definitions of "Trésor de la Langue Française informatisé"[1] (TLFi) (Dendien & Pierrel, 2003). We are interested in determining from a definition of TLFi the hypernymy relation that will be used for automatic construction of a thesaurus for image indexing and retrieval (Gheorghita, 2011). More precisely, we determine the possible hypernyms for a particular dictionary entry by exploring its definitions. In order to do this, we propose a weighting methodology that lets us assign to each lexeme the weight it has in a definition. This fact allows us to predict that the lexemes with the highest weight are the closest hypernyms of the defined lexeme in the dictionary.

2. Hypernymy in lexical models

Hypernymy is a lexical function that for a term *t* associates one or more other general terms. Logical definitions (or Aristotelian) are generally composed of a "genus" and "differentiae". In most of the definitions of this type, the hypernymy is represented by the relation "is-a". *A* is a hypernym of *B* if *B* is an *A* (a kind / type / kind of *A*) and if *A* is a classifier of *B*. This means that concept *B* is a specialization concept of *A*, and concept *A* is a generalization concept of *B*. For example, « mammal » is a generalization of « lion, wolf».

In linguistic resources like thesauri, WordNet, lexical entries are linked to other lexical entries by semantic relationships, so in WordNet the entry for *big* would somehow represent that its antonym is *small*. In this type of lexical model the relations that a word has to others partly determine the word's sense. In dictionaries, the meaning of lexemes is divided into several parts (Murphy, 2010). The information necessary for determining the semantic relations among words in dictionaries is contained in theirs definitions. Thus, we can determine the semantic relations between lexemes by a set of rules such as "*A* is the hyponym of *B* iff it has the same components as B, plus at least one more".

Compared to WordNet, the TLFi defines the meaning of a word only by a definition. The single information that can disambiguate the meaning of an input of TLFi is the domain[2] of definition. But only 31% of definitions have a domain. The definitions without a domain are assigned

[1] Treasury of the French Language Computerized

[2] There are a total of 7 786 domains

to the "generic" domain. It means that the sense of the word is also valid in the other domains. The majority of definitions of TLFi for nominal entries are logical where usually the first word of the definition is the hypernym of the entry. In the TLFi the semantic relations are not explicit. To determine the possible hypernyms of a TLFi entry, we calculate the weight of each noun in the definitions for a given domain. We assume that the nouns with the highest weight are the best hypernyms of the TLFi entry.

3. The word weighting method in the dictionary definitions

Our approach based on the analysis of the structure, the size and the meta-language of dictionary definitions, has allowed us to define a weighting method, which estimates the importance of lexemes in a definition. Thus, to calculate the final weight of the lexeme, we take into account the importance of the lexeme in a definition (local weighting), the importance of the lexeme in the collection of definitions for a given domain (overall weight) and the position of the lexeme in the chain of characters of the definition.

The importance increases proportionally to the number of times a word appears in the definition, and to the number of times a word appears in the collection of definitions for the given domain but is offset by the position in the definition.

The weight of a term *t* in a definition *d* for the domain *D* is defined as follows:

$$p_{t=} \frac{freq(t,d)}{\sum_i freq(t_i,d)} * \frac{N(d_t,c)}{N(d,c)} * \log_2 \frac{N_{pos}}{N_{pos}(t,ch)}$$

where:

$freq(t,d)$: frequency of a term *t* in the definition *d*

$\sum_i freq(t_i,d)$: frequency of all terms t_i in the definition *d*

$N(d_t,c)$: number of definitions in the collection for the domain *D* that contain the term *t*

$N(d,c)$: number of definitions in the collection for the domain *D*

N_{pos} : number of positions in the string of a definition *d*

$N_{pos}(t,ch)$: number of position of term *t* in the string *ch* of a definition *d*

The position of the lexeme is a very important indicator since the definitions of the dictionary are written by lexicographers according to some rules and using a specific meta-language. In the definitions, the meta-language terms occurred very often. Their weight is quite high compared with the weight of the other lexemes. It is for this reason that we created the specific classes for each type of meta-language terms. This fact allows us to distinguish the meta-language term from the lexeme and to increase or decrease the weight of the lexeme in dependence of its position with the meta-language term. Contrary to other weighting formulas as TF.IDF (Spark Jones, 1972) which favor the discriminants and rarest terms, our goal is to give more weight to the lexemes located at the beginning of the definition, considered as class representatives, and to the discriminant terms in the collection of definitions for a given domain, considered as specific characteristics.

According to the hypothesis made before, that the term with the higher weight is considered to be a best hypernym for the input *e* of TLFi, the weighting method is used to determine the list of possible hypernyms for the given term *e*. Jointly used with the inclusion model, which defines a set of rules of inheritance of properties from one class by a subclass, we build a thesaurus as a hierarchical tree where the terms are related by the relation “is-a”.

4. Evaluation of results and discussion

We applied our approach to 132 743 definitions that correspond to 51 778 nominal entries in a dictionary.

Table 1 shows examples of possible hypernyms for dictionary entries ranked by their weight in the definition. We noticed that the lexeme with the highest weight is not always the best hypernym of the dictionary entry. It is usually a meta-language term like *family of*, *form of* or a lexeme very characteristic of a given domain like *system* for *medical domain* and *tribunal* for *law domain*. This fact is explained by their high frequency in the collection of definitions for the given domain. However, the lexeme that can be considered as the best hypernym, like *fruit* for *avocado*, is in the list of the first three possible hypernyms. To determine it precisely, the frequent terms must be filtered and eliminated from the list of possible hypernyms.

We evaluated the quality of our methodology, by using the structured definitions of TLFi within the Definiens project (Barque, Nasr, & Polguère, 2010). In these definitions, the semantic markers are a central component (CC) and peripheral components (CP), which have been annotated manually. We assumed that the lexemes, with the highest weight in the list of possible hypernyms, must be located in the central component of the structured definitions. To prove our hypothesis, we calculated the precision. The precision of our results is the proportion of lexemes with the highest weight in the definitions determined as the central components in the structured definitions of Definiens project. Since the Definiens project has not been finished yet, we could only test our hypothesis for 15 000 dictionary entries.

Figure 1 shows the precision for the first three lexemes of maximum weight.

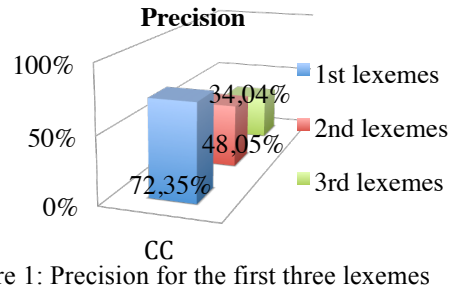

Figure 1: Precision for the first three lexemes