Is shallow parsing useful for unsupervised learning of semantic clusters?

Marie-Laure Reinberger, Walter Daelemans
CNTS - University of Antwerp - Belgium
{reinberg,daelem}@uia.ua.ac.be

Abstract. The context of this paper is the application of unsupervised Machine Learning techniques to building ontology extraction tools for Natural Language Processing. Our method relies on exploiting large amounts of linguistically annotated text, and on linguistic concepts such as selectional restrictions and co-composition.
We work with a corpus of medical texts in English. First we apply a shallow parser to the corpus to get subject-verb-object structures. We then extract verb-noun relations, and apply a clustering algorithm to them to build semantic classes of nouns. We have evaluated the adequacy of the clustering method when applied to a syntactically tagged corpus, and the relevance of the semantic content of the resulting clusters.

Keywords: semantics, knowledge representation, machine learning, text mining, ontology, selectional restrictions, co-composition.

1 Introduction

Semantic representations are useful for many natural language processing tasks, including information retrieval, word sense disambiguation, and automatic translation. However, in order to deal adequately with problems such as polysemy, these representations should be sufficiently rich and fine-grained. Today, the use of powerful and robust language processing tools such as shallow parsers allows us to parse large text collections and thereby provide potentially relevant information for extracting semantic knowledge. In order to decide what information is relevant for modeling semantic representations, we need strong linguistic hypotheses to guide the automatic extraction process. In this paper, we present a first step in an attempt to build tools for ontology extraction from scratch, on the basis of specific domain texts. At the same time, we intend to process as much as possible using strictly unsupervised methods on linguistically annotated texts.
We will present here in a first section our linguistic assumptions, followed by a description of the syntactic analysis we perform, a description of the semantic information extraction process, and an evaluation of our results.

2 Linguistic assumptions

Due to the richness and the diversity of the information that a word may carry, efficient lexical semantic representations should contain a multitude of informa-

tion of different kinds. In addition to the usual lexical information, these representations should include for example pragmatic information or knowledge of the world that might be useful to cope with problems such as ambiguity. In line with other data-oriented approaches to semantics, we start from the assumption that most of this information is present in plain texts in the way the words are organized and combined together to form complex expressions. The information that allows us to combine the right words together in order to produce meaningful expressions is assumed to be embedded in such texts, contained in the relations between the words of a complex expression. We take a broad perspective on these relations in that we do not restrict them to the hypernym/hyponym and meronymic relations, but that we also focus on information about the functionality of the concepts associated to a word (its uses, capacities etc.) that can be found in the way nouns and verbs, or nouns and adjectives are combined. The main problem lies then in finding a convenient way to get this information and retrieving it in an efficient way. We have chosen to begin this study by focusing on an unsupervised method, in order to gain insight in how far we could get (in terms of amount of and grain-size of the retrieved information) without human expertise.

An important assumption underlying our method is the hypothesis that syntax and semantics are not independent in natural language. They are closely related and interconnected, and we will refer here to this assumption as the principle of selectional restrictions: the syntactic structure of an expression provides relevant information about its semantic content.

The second hypothesis concerns the notion of co-composition [1]. Co-composition is an operation that occurs in the construction of meaning. If two elements compose an expression, each of them imposes semantic constraints on the other. In our studies, this is applied to the syntactic group noun-verb: the verb imposes restrictions on the noun, but the noun as well constrains the verb. In other words, each word in a noun-verb relation participates in building the meaning of the other word in this context ([2], [3]).

Related to these assumptions, we can then define two major tasks: (i) accessing the information, and (ii) organizing the information. Of course those tasks are related as the nature of the information retrieved will in some way influence its future organization. Our purpose is to build a repository of lexical semantic information, ensuring evolvability and adaptability. This repository can be considered as a complex semantic network. We could also label it an ontology, considering that an ontology is a collection of organized knowledge relative to a particular domain. An important point is that we assume that the method of extraction and the organization of this semantic information should depend not only on the available material, but also on the intended use of the knowledge structure. There are different ways of organizing it, depending on its future use and on the specificity of the domain. In this paper, we deal with such a specific domain, but one of our future objectives is to test our methods and tools on

different domains. This brings us to the choice, composition, and annotation of our corpus.

3 Syntactic analysis

We take a special interest in the compositional aspects of noun-verb relations. In order to provide information about these relations automatically in our corpus, we used the memory-based shallow parser which is being developed in Tilburg and Antwerp [4][1]. This shallow parser takes plain text as input, performs tokenization, POS tagging and phrase boundary detection, and finally finds grammatical relations such as subject-verb and object-verb relations, which are particularly useful for us. The software was developed to be efficient and robust enough to allow shallow parsing of large amounts of text from various domains.

In exploratory research, we used the Wall Street Journal corpus, but its vocabulary seemed not specific enough for our method, as we did not get enough occurrences for the different noun-verb pairs, at least for this first set of experiments on which we wanted to test the method. Consequently, we decided to test on texts representing more specific domains, and we used publicly available Medline abstracts, focusing on a particular medical subject. Our corpus is composed of the Medline abstracts retrieved by the Medline search engine under the queries "hepatitis A" and "hepatitis B". It contains about 4 million words. The shallow parser was used to provide a linguistic analysis of each sentence of this corpus, allowing us to retrieve semantic information of various kinds.

4 Semantic information extraction

Our method can be divided into two tasks. In a first step, we have used the syntactic information to perform a clustering of the nouns according to their relations with the verbs of the corpus. The second step will consist in building hierarchical relations between the clustered nouns, and between nouns and verb, making use of the results of the clustering.

4.1 Clustering

Method

As was mentioned earlier, the output of the shallow parser allows us to distinguish between noun-verb relations, where the noun appears as a subject in the expression, and noun-verb relations where it appears as an object. This lead us to focus particularly on the relation noun-verb and to use this information to

[1] See `http://ilk.kub.nl` for a demo version.

operate a clustering on the nouns according to the verbs they combine with[2].

Considering that most words have more than one meaning, we perform a *soft clustering*, in order to allow a word to belong to different clusters([5]) that represent different uses or meanings for this word.
The first step of the algorithm consists of processing the parsed text to retrieve the co-occurring noun-verb pairs, and remembering whether the noun appeared in a subject or in an object position. This step is performed with the use of a stoplist that skips all pairs implying the verbs to be or to have. We want to point out that we are not implying by doing so that those two verbs do not provide relevant information. They simply are too frequent and have such a broad meaning that we cannot, with this method and at this stage of the experiments, take them into account. We select then from the list we get the most frequent co-occurrences: the 100 most frequent noun-verb relations with the nouns appearing in the subject group, and the 100 most frequent relations where the noun is part of the object group. What we obtain is a list of verbs, each verb associated with a list of nouns that co-occur with it, either as subjects only or as objects only. Here is an extract of the list:

- acquiring_o: hepatitis infection virus disease
- associated_o: diseases cirrhosis DNA polymerase carcinoma HCC
- compensated_o: liver cirrhosis disease
- decompensated_o: liver cirrhosis disease
- decreased_s: rates prevalence serum incidence proportion number percentage
- estimated_s: prevalence rate virus incidence risk
- estimate_o: prevalence incidence risk number
- transmitted_o: hepatitis infection disease

The next step consists of clustering these classes of nouns according to their similarity. The similarity measure takes into account the number of common elements and the number of elements that differ between two classes of nouns. Each class is compared to all other classes of nouns. For each pair of classes C1-C2, the program counts the number of nouns common to both classes (sim), the number of nouns only present in C1 (dif1) and the number of nouns only present in C2 (dif2). If sim, dif1 and dif2 respect some predefined values the matching is considered to be possible. After the initial class has been compared to all other classes, all the possible matchings are compared and the one producing the largest new class is kept (in case of ties, the first one is kept). Each time a new cluster is created, the 2 classes involved are removed from the processed list. The whole process is iterated as long as at least one new matching occurs, resulting in the creation of a new cluster. We will describe the measures we used in the next section, along with the evaluation of the clustering.

[2] With *noun*, we refer to the head of an NP having a subject or object relation with the verb.

complete(o) starting(o)	contain(o)	develop(o) induce(o)	analyse(s) identify(s)	decrease(s) estimate(o)
immunization	antigen	hepatitis	aim	incidence
vaccine	virus	infection	objective	risk
vaccination	hepatitis	disease	purpose	proportion
	protein	cirrhosis	study	rate
	serum	carcinoma		

Table 1. Examples of extracted clusters

Results

We display in Table 1 some examples of steady clusters that appear in the results for each experiment in a series of experiments. Intuitively, the examples reported here seem to make sense, given the verbs they are associated to. For example, the nouns associated to the verbs *to decrease* and *to estimate* all name something that can be counted or represented by a number. The nouns associated to the verb *to complete* name something that can be fragmented or incomplete.
As we have used soft clustering, some words are associated to more than one cluster. This is the case for the word "hepatitis", e.g., which appears of course very often in this corpus. As shown in the table, "hepatitis" is associated with other diseases in the cluster of nouns representing nouns that can be combined with "to develop", and associated with other nouns representing things that can be considered as parts of a more important entity with the verb "to contain". But this anecdotal, intuitive approach does not tell us a lot about the general, objective, relevance of our clusters. Therefore, we need a method to measure this relevance, to ensure that the clusters indeed contain related words.

4.2 Evaluation of the clusters

We evaluate our clustering method at two different levels. The first level concerns the relevance of the clusters: do they associate semantically related words? The second level concerns the method itself: is the syntactic tagging really useful, or could we perform interesting clusters from unparsed text as well?

Relevance level

We evaluated the relevance of the clusters with the help of WordNet. Considering that we cannot automatically label the relations that unite the nouns of our clusters, we hoped that the variety of relations proposed by WordNet would fit the relations our clustering algorithm has built. The semantic information provided by WordNet is only used in the evaluation process. We do not intent to correct or enlarge the clusters with this information, as we wish to stay as much as possible within the paradigm of unsupervised learning.

	Number of clusters	Number of words	% of words clustered	Av. length of clusters
E1.1	120	153	94%	8.87
E1.2	28	105	64%	10.71
E2.1	155	148	91%	5.39
E2.2	32	108	66%	9.81

Table 2. Comparison of the percentage of words clustered and the average length of the clusters

	Number of WordNet pairs	Recall on the pairs	Number of incorrect pairs	Negative recall
E1.1	108	75%	11628	32%
E1.2	75	57%	5460	21%
E2.1	77	74%	10878	19%
E2.2	77	65%	5778	21%

Table 3. Recall and negative recall values for the different clustering experiments

We have extracted from WordNet all possible pairs consisting of two words present in the clusters, where the words of these pairs were linked in WordNet by a relation of synonymy, hypernymy, hyponymy or meronymy. The next step consisted of checking the presence of those pairs in the clusters. Of course, as the domain is very specific, not all the nouns present in the clusters are included in WordNet. Here are some examples of word pairs that could be extracted from WordNet:

hepatitis - disease (hypernymic relation)
blood - cells (meronymic relation)
aim - purpose (synonym)

Due to the fact that we are aiming at elaborating tools, we concentrate on experimenting and testing. Therefore, our clustering is not yet supposed to classify as many nouns as possible. As we have reduced the input to the clustering method to the 100 most frequent relations noun(subject)-verb and the 100 most frequent relations noun(object)-verb, the set of nouns was limited to 163 (some nouns appearing in the two sets). In the first experiment (E1.1), our criteria were that two classes of nouns could be merged if they had more than 2 common elements (sim>2), and not more than 5 different elements (dif<6). Once the clustering process ended, we considered as clusters the sets that contained at least 3 nouns. From the initial set of 163 nouns, 153 were clustered, which represents 94% of them (see Table 2).
We then fed WordNet with those 153 words, and we retrieved 108 pairs of words. As 27 of those pairs failed to appear in the clusters, we got, according to the

"WordNet sample" a recall of 75%. An evaluation of the precision score was difficult to settle as we do not have a gold standard of the "real" clusters. We therefore estimate a "negative recall", by generating incorrect pairs of words, and checking how many of them are present in the clusters. Those pairs are composed from non-related nouns, according to WordNet. We have generated about 11,000 pairs, of which about a third were present in the clusters, which in other words correponds to a negative recall of 32%. As the clustering is only a first step in an unsupervised ontology extraction process, it seemed sensible to focus on limiting the rate of errors and improve the results using other methods rather than investigating the mistakes. In order to improve the negative recall, we ran a new range of experiments (E2.1) where we allowed for more clusters to be formed (dif-sim<1). We kept the clusters containing two elements, but we eliminated the big clusters. A cluster was considered as too big when it contained more than 20 items, a number based on the biggest class associated to one verb. The same evaluation showed that about the same rate of words were clustered (91%). We obtained a good recall (74%), and a better negative recall (19%).
The elimination of the big clusters improves the precision score and is balanced by the creation of more small clusters, which improves recall. The weakness of both sets of results lies in the high number of clusters produced: 120 clusters for the first experiment, and 155 clusters for the second.
We tried to reduce the number of clusters by removing the smaller ones from both sets of previous results (experiments E1.2 and E2.2). We obtained for E1.2 a group of 28 clusters, which corresponds to 64% of the words, with a negative recall of 19%, but a recall of only 57%. The results for E2.2 were quite similar with a better recall of 65%. We conclude from this that relevant information can be found in the small-sized clusters, and that by removing the small clusters, we lose this information without improving the negative recall measure.
The experiment that rates the best score according to our objectives is experiment E2.1. It gives us the lowest negative recall, a good recall, and a high rate of the set of initial words are clustered. Its weak point is the numerous clusters generated. But this clustering is only the first result in an ontology extraction tools process, and the next steps will aim at improving the results of the clustering and making the clusters more precise.
A summary of the results discussed above appears in tables 2 and 3.

Efficiency level

The second step of our evaluation consisted in comparing the results of the clustering algorithm on parsed text with the results we would get processing on plain text, in order to get a baseline. Our hypothesis is that the clustering performed on a syntactically analyzed text is more accurate than one performed on raw text. But we are also interested in the magnitude of the difference in performance between both methods: is it really worth the trouble to analyze the corpus syntactically, or can we get useful results already with raw text, results

that we could then improve by retrieving more semantic information from more text?

	Nb of clusters	Nb of words	Nb of words clustered	% of words clustered
E2.1	155	163	148	91%
4000 m.f.bg	38	1663	206	12%
5000 m.f.bg	52	1931	263	14%

Table 4. Percentage of words clustered using parsed text and using plain text

	Nb of correct pairs	Recall	Nb of negative pairs	Negative recall
E2.1	77	74%	10878	19%
4000 m.f.bg	51	29%	21037	6%
5000 m.f.bg	75	27%	34641	5%

Table 5. Recall and negative recall values for the clustering on parsed text and on plain text

We ran a set of experiments on plain text, using bi-grams as the equivalent for plain text of the noun-verb pairs in the annotated text. We have compared the two methods on the basis of the number of words clustered. The clustering on annotated text worked on 163 words corresponding to the 200 most frequent relations, of which 148 were clustered in experiment E2.1. In the bigram experiment, it appeared that considering the 4000 most frequent bigrams corresponded to 1663 words and that 206 of those words were clustered at the end of the process. We ran the clustering algorithm on different numbers of bigrams, and the results were quite similar. As shown in Table 4, and considering that the bigrams, even with the use of a stoplist, select all kinds of words, the percentage of words contained in the clusters was very low, which means that a lot of words have to be taken into account to cluster only a (comparatively) small number of nouns. As expected, the recall on the clusters using the WordNet pairs was low, and never reaching more than 30%. The best measure we obtained for all bigram sets was the negative recall, which never went over 6%. We give the results in Table 5 for the 4000 and the 5000 most frequent bigrams. We can see there that a difference of 1000 bigrams does not change significantly the recall values.
The results we get show that the use of annotated text improves the rate of words clustered and the recall. The difference of those two rates is important enough to balance the better negative recall, and to let us consider that performing a syntactic analysis prior to the clustering is useful.

5 Ongoing work: Labeling and building a hierarchy

The next task in our project consists of labeling the relations between the nouns and building a hierarchy. To further pursue an unsupervised approach, the semantic labeling should be done automatically. We therefore do not intent to use WordNet and the different relations it proposes, but will try to get those semantic relations directly from the corpus.
The clustering we are performing does not provide any information concerning the kind of relations between the clusters, hence between the words. However, the different elements of a cluster have something in common that can be specified as a relation. We can focus on two types of information: the relations between words belonging to the same cluster, and the relations between the clusters of nouns and the verbs associated to them according to the relation verb-noun (subject or object). We are planning to perform this by using methods involving pattern matching or association rules ([6]), and automatic methods for constructing hierarchies ([7], [8]).

6 Conclusions and perspectives

We have shown that unsupervised learning methods can be used to retrieve semantic information from text when a shallow syntactic analysis is available. This syntactic analysis proved to be useful as the clustering performed on the parsed text gave better results than the one performed on plain text. The next step of this research is to elaborate the conceptual knowledge sets for the clusters of nouns. Another interesting extension would consist in considering the groups of verbs associated to the clusters of nouns. That information could allow us to cluster the verbs and get selectional preferences associated with classes of verbs, but also to relate nouns to verbs, where these relations represent the semantic functions of the concept associated with the noun. Yet another issue is the retrieval of the prepositions that introduce a nominal complement and use this information to make the information associated with nouns more specific.

7 Acknowledgments

This research was carried out in the context of the OntoBasis project, sponsored by IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders).

References

1. Pustejovsky, J.: The Generative Lexicon. MIT Press (1995)
2. Gamallo, P., Agustini, A., Lopes, G.P.: Selection restrictions acquisition from corpora. In: Proceedings EPIA-01, Springer-Verlag (2001)

3. Gamallo, P., Agustini, A., Lopes, G.P.: Using co-composition for acquiring syntactic and semantic subcategorisation. In: Proceedings of the Workshop SIGLEX-02 (ACL-02). (2002)
4. Daelemans, W., Buchholz, S., Veenstra, J.: Memory-based shallow parsing. In: Proceedings of CoNLL-99. (1999)
5. Faure, D., Nédellec, C.: Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In: Proceedings EKAW-99. (1999)
6. Maedche, A., Staab, S.: Semi-automatic engineering of ontologies from text. In: Proceedings of SEKE-00. (2000)
7. Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings ACL-99. (1999)
8. Berland, M., Charniak, E.: Finfing parts in very large corpora. In: Proceedings ACL-99. (1999)
9. Agirre, E., Martinez, D.: Learning class-to-class selectional preferences. In: Proceedings CoNLL-01. (2001)
10. Caraballo, S.A., Charniak, E.: Determining the specificity of nouns from text. In: Proceedings SIGDAT-99. (1999)
11. Gamallo, P., Gasperin, C., Agustini, A., Lopes, G.P.: Syntactic-based methods for measuring word similarity. In: Proceedings TSD-01, Springer-Verlag (2001)
12. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems **16** (2001)
13. McCarthy, D., Carroll, J., Preiss, J.: Disambiguating noun and verb senses using automatically acquired selectional preferences. SENSEVAL-2 (2001)
14. Wagner, A., Mastropietro, M.: Collecting and employing selectional restrictions. Technical report (1996)