Unsupervised Text Mining for Ontology Extraction: An Evaluation of Statistical Measures

Marie-Laure Reinberger, Walter Daelemans

CNTS/University of Antwerp
Universiteitsplein 1
2610 Wilrijk - BELGIUM
{marielaure.reinberger,walter.daelemans}@ua.ac.be

Abstract

We report on a comparative evaluation carried out in the field of unsupervised text mining. We have worked on a parsed medical corpus, on which we have used different statistical measures. Using those measures, we rate the verb-object dependencies and we select the most reliable ones according to each measure. We then apply pattern matching and clustering algorithms to the classes of dependencies in order to build sets of semantically related words and establish semantic links between them. Finally, we evaluate the impact of the statistical measures used for the initial selection of the dependencies on the quality of the results.

1. Introduction

Today, the use of powerful and robust language processing tools allows us to parse large text collections and thereby provide potentially relevant information for extracting semantic knowledge. Statistics can help us selecting the relevant information for modeling semantic representations, but choosing the right statistical tool is a crucial point.

The context of this experiment is the project OntoBasis [1]. One purpose in this project is the development of linguistic tools for unsupervised ontology extraction. Our general approach consists of working with domain specific corpora, on which we apply a syntactic parsing. Then, we select specific syntactic structures, on which we perform clustering and pattern matching in order to extract semantic relations between nominal expressions.

The study we are reporting on here focuses on the selection of syntactic structures, more precisely on a comparison of various statistical methods that allow us to rate the goodness of structures detected in the corpus, in an unsupervised way. We have opted for extraction techniques based on unsupervised learning methods (Reinberger et al., 2003) since these do not require specific external domain knowledge such as thesauri and/or tagged corpora.

We rely on the principle of selectional restrictions, that states that syntactic structures provide relevant information about semantic content, in that case that heads of object phrases co-occurring with the same verb share a semantic feature. We rely also on the notion of co-composition (Pustejovsky, 1995). If two elements are composed into an expression, each of them imposes semantic constraints on the other, here in consequence each word in a noun-verb relation participates in building the meaning of the other word in this context (Gamallo et al., 2001; Gamallo et al., 2002).

2. Experimentation

We have worked with a 5M words corpus composed of Medline abstracts related to the hepatitis disease. In a specific domain, an important quantity of semantic information is carried by the noun phrases (NP). At the same time, the NP-verb relations provide relevant information about the NPs, due to the semantic restrictions they impose. Therefore, we applied to this corpus a memory based shallow parser that detects subject-verb-object structures (Buchholz et al., 1999; Buchholz, 2002; Daelemans et al., 1999) [2]. This shallow parser gives us the possibility to exploit the verb-object dependencies. The selectional restrictions associated with this structure imply that the NPs co-occurring, as the head of the object, with a common set of verbs, share semantic information. This semantic information can be labeled as "functional", due to the semantic role of the verb.

Our corpus provides us with a huge number of those syntactic structures associating a verb to a nominal string (NS), but we have to deal with the fact that the parser produces also some mistakes (f-score for objects is 80 to 90%), and that not all verb-object structures are statistically relevant. Therefore, we need to find a way to select the most reliable dependencies, before applying to them automatic techniques for the extraction of ontological relations. This step can be achieved with the help of statistical measures that take into account the frequency (f) and probability (P) of occurrence of the different elements of the syntactic structure. But there is a wide range of measures, requiring more or less computing time. And there is a priori no indication that one measure would perform better than another on our data.

Therefore, we carried out an evaluation, using 5 different measures that put the stress on different aspects of the syntactic structures:

- a simple frequency measure that we will consider as a baseline: $Fnv = f(n,v)/(f(n) + f(v))$
- a measure based on the probability of appearance of the verb-object dependency: $Pnv = f(n,v)/(f(v)$
- the Hindle (Hindle and Rooth, 1993) mutual information measure (using occurrence probabilities P), which put the stress on the strength of the verb-object relation: $Hnv = log\{P(n|v)/[P(n) * P(v)]\}$

[1] See `http://wise.vub.ac.be/ontobasis/`

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

with:

$$P(n|v) = f(n,v)/\sum_{v'} f(v')$$
$$P(v) = f(v)/\sum_{v'} f(v')$$
$$P(n) = f(n)/\sum_{n'} f(n')$$

- the Resnik (Resnik, 1997) measure, which computes for each verb its selectional preference strength Sr(v); this measure is high when the NSs that combine with the verb as objects are infrequent:

  $Rnv = P(n|v) * Sr(v)$

  with $Sr(v) = \sum_n \{P(n|v) * \log[P(n|v)/P(n)]\}$

- the Jaccard measure, which considers the number of contexts (#ctxt) in which a NS (n) appears:

  $Jnv = \log_2 P(n|v) * \{\log_2 [f(n)/\#ctxt(n)]\}$

Each measure has been computed with 2 different values for f(n), considering:

- only the occurrences of n in verb-object structures;

- all occurrences of n in the corpus;

For each measure, we have selected the best 500, 700 and 900 associations [class of verbs: NS]. Each of them is composed of a set of verbs associated to the NSs they frequently occur with according to the measure concerned:

Here are some examples of such associations:

- consume drink abuse: alcohol

- combat terminate: chronic_hbv_infection

At this point, we need a method to gather NSs according to their common semantic features. Clustering only requires a minimal amount of "manual semantic preprocessing" by the user, and clustering on NSs can be performed by using different syntactic contexts, for example noun+modifier relations (Caraballo and Charniak, 1999) or dependency triples (Lin, 1998).

As our intention is to put the stress on the selection of the structures, we have chosen to apply on those sets of associations a naive clustering method based on the similarity between two classes of verbs. This similarity depends on the number of elements common to 2 classes and of the statistical scores of the verbs.

As each class of verbs is associated to a nominal string, this clustering will build at the same time classes of NSs, with an NS only belonging to one cluster. By performing this clustering, we mean to exploit the functional relation that occurs between a verb and its direct object. During the first pass, NSs will be joined two by two. In the next passes, the sets of NSs are joined two by two.

The sets of NSs gathered by the clustering algorithm share a functional information, for example:

- face_mask protective_eyewear mask glove

- transcriptase transcriptase_inhibitor transcription-polymerase_chain_reaction

Pattern matching (Berland and Charniak, 1999) has proved to be a efficient way to extract semantic relations, but one drawback is that it involves the predefined choice of the semantic relations that will be extracted. Here, we will combine it with the results of the clustering. Therefore, the last step of this experiment consists of creating links between the sets of NSs. We have retrieved all the patterns [NS1-preposition-NS2] in our corpus. A set of NS2 is formed according to the fact that each appear with the same couple NS1-preposition. We have as a result a list of elements: [NS-preposition-set of NS]. Then, we check for similarities between those sets of NS and the clusters obtained previously. In case of similarity (common NS in both sets), the cluster is increased with the new elements and the link labeled by the preposition is added. The last step consists in checking if some more clustering is possible among the resulting elements [NS-preposition-set of NS] (increased by the content of the clusters). We give below two examples of the final structures:

- [recurrence transmision] of [infection hepatitis_B_virus viral_infection HCV hepatitis_B HCV_infection disease HBV HBV_infection viral_hepatitis]

- [heparin blood_pressure blood blood_loss] during [aortic_surgery operation apostosis surgery coronary_angiography hemipathectomy coronary_artery_bypass emergency-surgery cardiac_surgery surgical_resection hemodialysis procedure dialysis transplantation]

3. Evaluation

As we deal with medical data, we perform an evaluation of the classes and clusters we obtain with UMLS (Unified Medical Language System). The evaluation of extracted clusters is problematic, as we do not have any reference or model for the clusters that we want to build. At the same time, we want this evaluation to be automatic.

Considering as a reference the set of NS that appear in the clusters, we retrieve from UMLS all pairs formed with two NS from the reference set and sharing a semantic relation in UMLS. Then, we check how many of those pairs appear at least in one of the clusters. Using this information, we compute a recall value R, a precision value P, and a classic F-measure:

$$\frac{1}{\frac{1}{2P} + \frac{1}{2R}}$$

We want to point out the fact that we cannot evaluate exhaustively the content of our clusters, as some of the NS they contain are unknown in UMLS. This evaluation must therefore be considered as a partial evaluation.

4. Results

After the clustering step (Figure 1 and 2), the results are more contrasteddivergent in the pool where all occurences of NS were considered (Figure 2), and it is the Jaccard measure that gets the best results (F-measure 0.25-0.35). This may be due to the fact that this measure takes into account

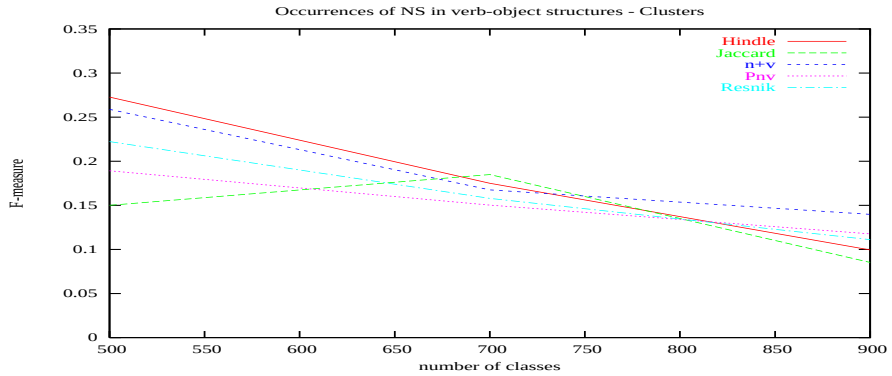

Figure 1: *F-measure after the clustering, considering only the occurrences of NS in verb-object structures*

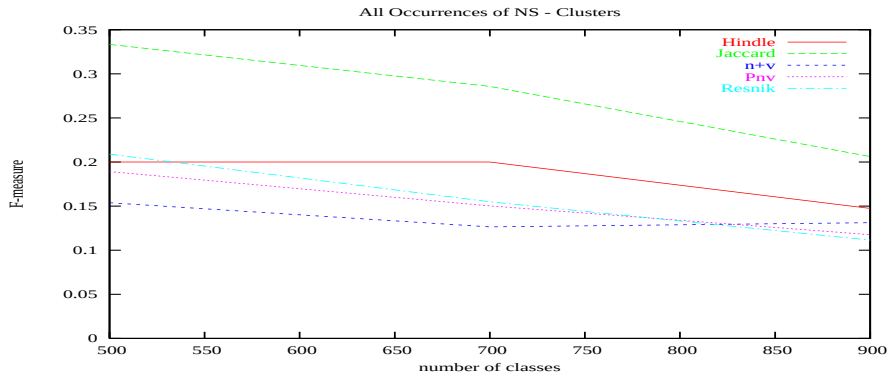

Figure 2: *F-measure after the clustering, considering all occurrences of NS*

the context of appearance of the NS. The Hindle measure gets the second best score (0.2).

But after the next step of the process (Figure 3 and 4), when we have increased the clusters and established some links by using the information extracted with the pattern [NS-preposition-NS], the situation changes. First, in the experiment where only the NS occurring in verb-object structures were counted (Figure 3), the results are globally worse. One reason is that the prepositional pattern is not selecting the NS highly rated in verb-object structures.

If we consider the part of the experiment where only occurrences of NS in verb-object structures were considered (Figure 1 and 3), it appears that the baseline measure performs as well as more elaborated measures.

Then, if we consider the part of the experiment where all occurrences of NS were considered (Figure 2 and 4), we notice that the Hindle measure performance has increased (0.32 for 500 classes, 0.12 for 900 classes) while Jaccard measure has decreased (0.12 for 500 classes, 0.05 for 900 classes). That is due to the initial selection of verb-object dependencies: the Hindle selection allowed many prepositional patterns to be added to the clusters. On the other hand, the Jaccard measure tends to select dependencies containing elements that do not appear as frequently in a prepositional pattern. As a consequence, the clusters produced using Hindle measure combine better with the patterns than the clusters using Jaccard measures, hence an improvement or a decrease of the performance. In most cases, the baseline measure (frequency measure) gets the worst results. The Resnik measure performs poorly also, just above the baseline, which shows that the selectional preference strength of verbs does not constitute relevant information for this task.

In a recent study concerning the automatic acquisition of taxonomies (Cimiano et al., 2003), the results of different statistical measures in a concept classification task are compared considering a measure threshold. We consider that this might induce a bias as, for the same threshold, two measures could select a very different number of terms . For that reason, we have carried out our comparison considering the initial number of verb-object classes selected.

5. Conclusion

This study has shown that depending on the data we perform the clustering on, different measures will produce different results. Whether we consider all occurrences of the NS or not induces an important discrepancy for some F-measures. Adding information (here, through the prepositional pattern) improves the results in some cases (Hindle), but might lower them for some measures (Jaccard).

However, in one experiment, the baseline measure levels with other more elaborated measures, which proves that for some tasks, a simple frequency measure can provide good results.

Therefore, the choice of the statistical measure, hence the adaptation of the measure to the data is a crucial point

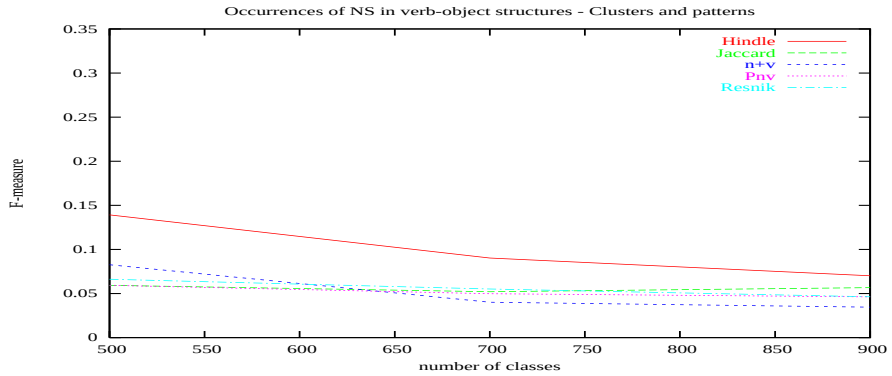

Figure 3: *F-measure after the pattern matching, considering only the occurrences of NS in verb-object structures*

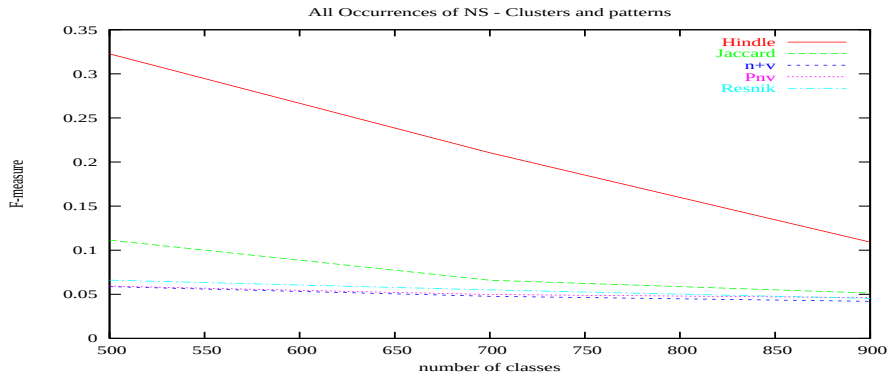

Figure 4: *F-measure after the pattern matching, considering all occurrences of NS*

for the selection of dependencies for the modeling of semantic representations. Unfortunately, the optimal measure for a particular corpus has to be determined experimentally, and cannot be decided in advance.

Acknowledgements This research was carried out in the context of the OntoBasis project (GBOU 2001 #10069), sponsored by the IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders). Our academic partners in this project are STARLab and WISE at VUB (Free University of Brussel).

6. References

Berland, Matthew and Eugene Charniak, 1999. Finding parts in very large corpora. In *Proceedings ACL-99*.

Buchholz, Sabine, 2002. Memory-based grammatical relation finding. In *Proceedings of the Joint SIGDAT Conference EMNLP/VLC*.

Buchholz, Sabine, Jorn Veenstra, and Walter Daelemans, 1999. Cascaded grammatical relation assignment. In *Proceedings of EMNLP/VLC-99*.

Caraballo, Sharon A. and Eugene Charniak, 1999. Determining the specificity of nouns from text. In *Proceedings SIGDAT-99*.

Cimiano, P., S.Staab, and J.Tane, 2003. Automatic acquisition of taxonomies from text: Fca meets nlp. In *Proceedings ATEM03*.

Daelemans, Walter, Sabine Buchholz, and Jorn Veenstra, 1999. Memory-based shallow parsing. In *Proceedings of CoNLL-99*.

Gamallo, Pablo, Alexandre Agustini, and Gabriel P. Lopes, 2001. Selection restrictions acquisition from corpora. In *Proceedings EPIA-01*. Springer-Verlag.

Gamallo, Pablo, Alexandre Agustini, and Gabriel P. Lopes, 2002. Using co-composition for acquiring syntactic and semantic subcategorisation. In *Proceedings of the Workshop SIGLEX-02 (ACL-02)*.

Hindle, D. and M. Rooth, 1993. Structural ambiguity and lexical relations. *Computational Linguistics,*, 19.

Lin, Dekang, 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL-98*.

Pustejovsky, James, 1995. *The Generative Lexicon*. MIT Press.

Reinberger, Marie-Laure and Walter Daelemans, 2003. Is shallow parsing useful for the unsupervised learning of semantic clusters? In *Proceedings CICLing03*. Springer-Verlag.

Reinberger, Marie-Laure, Peter Spyns, Walter Daelemans, and Robert Meersman, 2003. Mining for lexons: applying unsupervised learning methods to create ontology bases. In *Proceedings ODBASE03*. Springer-Verlag.

Resnik, P., 1997. Selectional preferences and sense disambiguation. In *Proceedings ACL-SIGLEX97*.