Automatically selecting domain markers for terminology extraction

Jorge Vivaldi¹, Horacio Rodríguez²

¹Institute for Applied Linguistics, Universitat Pompeu Fabra
La Rambla 30-32, 08002 Barcelona, Spain
jorge.vivaldi@upf.edu
²Software Department, Universitat Politècnica de Catalunya
c/ Jordi Girona 31, 08034 Barcelona, Spain
horacio@lsi.upc.es

Abstract

Some approaches to automatic terminology extraction from corpora imply the use of existing semantic resources for guiding the detection of terms. Most of these systems exploit specialised resources, like UMLS in the medical domain, while a few try to take profit from general-purpose semantic resources, like EuroWordNet (EWN).

As the term extraction task is clearly domain depending, in the case a general-purpose resource without specific domain information is used, we need a way of attaching domain information to the units of the resource. For big resources it is desirable that this semantic enrichment could be carried out automatically.

Given a specific domain, our proposal aims to detect in EWN those units that can be considered as domain markers (DM). We can define a DM as an EWN entry whose attached strings belong to the domain, as well as the variants of all its descendents through the hyponymy relation. The procedure we propose in this paper is fully automatic and, a priori, domain-independent. The only external knowledge it uses is a set of terms, which is an external vocabulary, which is considered to have at least one sense belonging to the domain.

1 Introduction

Some approaches to automatic terminology extraction from corpora imply the use of existing semantic resources for guiding the detection of terms. Most of these systems exploit specialised resources, like UMLS¹ in the medical domain, while a few try to take profit from general-purpose semantic resources, like EWN².

As the term extraction task is clearly domain depending, in the case a general-purpose resource (e.g. an ontology) without specific domain information is used, we need a way of attaching domain information to the units of the resource. This semantic enrichment can be carried out manually, but, for big resources, the cost of manually examining the whole data set in order to look for items belonging to the specific domain makes desirable an automatic, or at least a semi-automatic, procedure.

Given a specific domain, our proposal aims to detect in EWN, a wide-coverage general-purpose lexico-semantic ontology, those units that can be considered as domain markers (DM). We can define a DM as an EWN entry (a synset) whose attached strings belong to the domain, as well as the variants of all its descendents through the hyponymy relation. The procedure we propose in this paper is fully automatic and, a priori, domain-independent. The only external knowledge it uses is a set of terms, which is an external vocabulary, which is considered to have at least one sense belonging to the domain.

The domain of Medicine has been selected because our previous experience in this area, its relatively large coverage in EWN, the existence of other works in this domain, and the availability of public-domain vocabularies.

After this introduction, section 2 briefly discusses some related approaches, then section 3 presents an overall description of our proposal. Two empirical evaluation procedures have been developed: direct and indirect one.

Both are presented in section 4. Finally, in section 5, some conclusions and lines of future work are stated.

2 Related approaches

(Magnini, Cavaglià, 2000) have enriched WN with domain information. Such task has been done on the basis of a general classification that includes 164 domains/subdomains (structured in a rather flat taxonomy). Following a semiautomatic procedure, one or more domain tags has been assigned to each synset.

In an automatic term extraction system, applied to the medical domain, (Vivaldi, Rodríguez, 2002) use Medical Borders, i.e. synsets in EWN for which it is assumed that they belong to the medical domain and also all their hyponyms do. About 50 medical borders were manually identified and used as a basis for term extraction.

(Montoyo et al, 2001) propose a way of enriching WN with about 30 IPTC³ subject codes. Their approach follows the Specification Marks Method, previously used for Word Sense Disambiguation tasks. Also (Buitelaar, Sacaleanu, 2001) propose a method for domain specific sense assignement using GermaNet (a resource similar to WN) together with relevance measures. A closely related task is the automatic extraction of domain ontologies from general ones using domain corpora. (Missikoff et al, 2002) present an interesting approach.

3 Description of the system

For our purposes, we consider only the nominal part of EWN (WNn) and the hyperonymy/hyponymy relations. A synset s is considered as a DM of a domain if in the set of s and its descendents the density of synsets belonging to the domain D is over a predefined threshold. The core of our system is to select a set of DM candidates, to define the likelihood of each candidate (domainhood), and to accept as true DM those over a threshold. However, we must take into consideration that being a DM can be

¹ http://umlsinfo.nlm.nih.gov
² http://www.illc.uva.nl/EuroWordNet/
³ http://www.iptc.org
considered not an absolute property but a probability or likelihood of belonging to the domain.

The way of selecting DM consists of locating zones in WNn where the estimated density of synsets belonging to D is over the threshold. For measuring such density we have used as external knowledge source a vocabulary (V0) of terms that, with high confidence, are considered to belong to D. In our experiments, we have used as well a validation corpus. Using this additional knowledge source, if available, leads to an improvement of the results.

Our system proceeds in two steps:

1. An initial set of DM, DMinic, is built following these ideas and using V0 as Knowledge Source. Results of using DMinic are referred as Automatic DM in section 4.

2. If a validation corpus is available, a second set is derived starting from a state associated to the DMinic and looking for a better solution through a greedy search on the neighbourhood of the synsets belonging to DMinic. 1. Results of using this set are referred as Automatic & Improved DM in section 4.

Two different procedures to calculate which synsets may be considered as DM have been developed. The first one splits the set of synsets attached to words in V0 into four classes according the number of senses related to such words. A probability of being a DM has been attached to each one of them. Following this calculation it looks in the EWN hypernymy chain for a given probability threshold (i.e. the system tries to select in the hypernymy chain the stop point for placing the DM).

We will consider that a zone Zs in WNn is the subtree rooted at s taken into account hyponymy relations. We can model l as DM as the probability that a randomly selected synset belonging to Zs belongs as well to D. For doing so, we will split Zs into three sets C1, C2 and C3:

C1(s) = {x ∈ Zs | x belongs to the first class}
C2(s) = {x ∈ Zs | x belongs to other classes}
C3(s) = Zs - C1(s) - C2(s)

We will use a random variable S ranging on Booleans. We will associate as well random variables C1, C2 and C3 for modeling belonging to the corresponding sets:

C1(s) = true if x ∈ C1 and false otherwise for i = 1 to 3
So, we can write P(S(s) = yes|s) as the probability, given s, of belonging to D, and, simplifying the notation, P(S|s). In a similar way, we can write P(C1|s) instead of P(C1|s) = yes|s) and the same for C2 and C3. Being {C1, C2, C3} a partition of Zs, the conditional probability can be written as:

P(S|s) = P(S|C1) ⋅ P(C1|s) + P(S|C2) ⋅ P(C2|s) + P(S|C3) ⋅ P(C3|s)

We will make the simplification assumption that P(S(s)|C1(s)) does not depend on the particular s. Applying this assumption and normalising we can write the expression as:

P(S|s) = P(C1|s) + α ⋅ P(C2|s) + β ⋅ P(C3|s)

In this formula, all the terms P(Ci|s) can be easily computed using MLE from a training corpus and α and β are parameters of our model. We have experimented with several values of α and β in a development corpus for getting the best values. Given the origin of Ci, C2 and C3, it is clear that α < 1 and β << α.

Let D designate a domain. Our method consists on the following steps:

1. Select V0. We will consider all the members of V0 as belonging to D. It will be assumed as well that every w ∈ V0 has at least one sense in WNn that belongs to D.

2. Remove from V0 all its members not covered by WNn. Let V0 designate this new set.

3. From V0 we build SYN, i.e. the union of SYNw for all w in V0.

4. For every s belonging to SYN so that it contains only one variant, being this variant monosemic, we compute its hypernymy chain (in fact more than one chain could be followed from one synset due to the possibility of having more than one hypernym) until reaching a top of the hierarchy (in the case of the medical domain, reported in section 4, all the 11 tops of WNn are reached, for other domains may be not all the tops are reached).

5. For all the synsets s, belonging to any of the chains obtained in 4) we compute its scoring l(s).

6. For all the chains obtained in 4) a break point has to be determined. Conflicts can be produced between chains having a common suffix, but, due to the way of computing l(s), when more than one chain reach a synset its scoring reflects the likelihood of all the descendents and, so, the number of conflicts is small and can be solved with local heuristic rules.

Obviously, the higher likelihood in each chain is found for the terminal synsets, i.e. the origin of the chain, while the lowest likelihood corresponds to the tops. However, the shape of the figures is not uniform and two different behaviors usually occurs. In the first one likelihood falls monotonically as we climb on the hierarchy, in the other one or more local maxima occur. Our algorithm focuses on this late case. Figure 1 presents the result of l(s) for the chain extending the term ‘tachycardia’. This presents the typical shape of a chain without maximum.

The second procedure for calculating the DM just takes into consideration those entries in V0 that are monosemic for performing, starting with them, a best first search of those synset that may work as DM. This procedure did not produced any improvement in our evaluation.

The second step of our approach establishes that a given list of DM is a state. Three types of primitive operations can be applied on a given state for allowing the transition to a new state: removing one of the member of the current DM, climbing up in the hierarchy substituting one of the

---

4 Choosing a terminal synset s as DM is not useful because only this synset belongs to Zs.
members of DM by one of its direct hypernyms or moving down in the hierarchy substituting one of the members of DM by the whole list of its direct hypernyms.

4 Evaluation
Our empirical evaluation schema includes two different steps: a direct evaluation and an indirect one. Direct evaluation consists on directly comparing the set of DM produced by our methods with two other approaches. The indirect evaluation consists on using the results in a terminology extraction task (see Vivaldi, 2001).

Direct evaluation: Selecting domain markers
The vocabulary $V_D$ was obtained from MedicineNet\(^5\). A test with another medical resource, with more terms (65,534) but also more noise\(^6\), produced worse results. Our method is sensitive to the quality of the data in the $V_D$, due to the assumptions stated in Section 3.

MedicineNet contains 11,514 medical terms from which only 2,487 exist in WN. So, $|V_D| = 11,514$ and $|V'_D| = 2,487$. Included in SYN there are 571 monosemic synsets that are candidates to be DM. Three values of $\alpha$ have been tested. $\beta$ has been set to $\alpha/100$ in all the experiments.

Table 1 presents the six highest scored synsets using the $M_{\text{max}}$ method, with $\alpha = 0.5$. $h_1$, $h_2$ and $h_3$ refer respectively to the number of hyponyms, the number of hypernyms with at least one medical sense and the number of hypernyms with just one medical sense. Only the variants represented in $V_D$ have been included in the table.

| synset          | score | $h_1$ | $h_2$ | $h_3$ | variant               |
|-----------------|-------|-------|-------|-------|----------------------|
| 08648329        | 0.63  | 14    | 13    | 6     | malignancy           |
| 08647140        | 0.57  | 27    | 21    | 11    | --                   |
| 08603909        | 0.57  | 14    | 12    | 5     | cardiovascular disease |
| 08693652        | 0.48  | 23    | 13    | 10    | anxiety disorder     |
| 08636825        | 0.45  | 0     | 0     | 0     | pathology            |
| 03729776        | 0.42  | 37    | 22    | 10    | hormone              |

Table 2 presents the overall results for different parameters showing the intersections of $O$, $VR$ and $MC$, where $O$ is the method introduced in this paper, $MC$ is the method presented by Magnini et al. (2000) and $VR$ is the method proposed by Vivaldi (2001).

| Method   | $\alpha$ | #DM | 101 | 110 | 111 | 001 | 010 | 100 | 011 |
|----------|----------|-----|-----|-----|-----|-----|-----|-----|-----|
| Mthreshold | 0.5      | 246 | 2267| 78  | 954 | 1896| 361 | 1463| 865 |
| Mmax     | 0.5      | 66  | 787 | 20  | 2434| 462 | 395 | 521 | 2299|
| Mdelta   | 0.5      | 92  | 1775| 45  | 1446| 1158| 356 | 496 | 1603|
| Mthreshold | 0.1     | 358 | 2486| 67  | 735 | 1526| 279 | 474 | 1235|
| Mthreshold | 0.01    | 374 | 2511| 67  | 710 | 1139| 279 | 474 | 1622|

For MC experiments, 9 tags have been selected as belonging to the medical domain: medicine, dentistry, pharmacy, radiology, surgery, physiology, etc. The total amount of synsets having a medical tag was 5,073. Because $VR$ experiments were performed on WN1.5 and $MC$ on WN1.6, we have considered only the synsets having a direct mapping from WN1.6 to WN1.5. So, the total result is reduced to 3,762.

For $VR$ experiments, 58 DM were manually selected. The number of synsets under these DM was 5,982.

For $O$ (mainly if the user do not have some knowledge of the domain and/or familiarity with ontologies). What we attempt to do here is to apply the same term extraction methodology but using as $DM$ the domain markers defined in this paper. We tested the behavior of the

\(^5\)http://www.medicinenet.com/

\(^6\)Such noise is due to words not belonging to the domain.
proposed methodology using two documents\(^7\). Such documents have been linguistically processed as usual in most of the NLP tasks. We evaluate the results using the standard measures of precision and recall. For testing the performance of the improved procedure we perform several tests in the training corpus using different set of parameters. Finally we choose one of the resulting sets of improved DM taking into consideration both the precision score\(^8\) and the precision value (for 30\% of recall). We apply the resulting set of DM to extract medical terms in the test corpus. Figure 2 shows the results obtained with the training corpus using the automatic method of selecting the DM. It shows that there is a fall in recall for intermediate figures of the precision. This loss may or may not be relevant; it fully depends on the usage of the extracted terms. In any case, this is the price we have to pay for reducing the cost of manually selecting DM.

Figure 2. Evaluation of the DM in test corpus

Figure 3. Evaluation of the DM in training corpus

The results obtained using the improved set of automatic DM are shown in Figure 3. There is an improvement of the results, even against the manual DM for some values of recall. Also there is minor but steady enhancement of the improved automatic procedure against the automatic procedure; although such improvement does not seem to be statistically significant.

5 Conclusions

This paper shows how public available vocabularies may be used to enrich general-purpose resources with domain information in a fully automatic way. For such a purpose, we have defined a likelihood estimate that has been tested using different parameters. Also, we found a method for further refining this calculation. We have successfully tested the list of domain markers comparing it with other approaches. Moreover, we have obtained relevant results in extracting medical terminology from a specialised corpus. A possible way of taking profit of the two approaches could be using the automatic DM as an initial step followed by a manual one. This possibility will be explored in a near future. We also foresee to check the performance of the proposed method in areas different from Medicine.

6 References

Bourigault, D., C. Jacquemin and MC. L’Homme (eds), (2001). Recent Advances in Computational Terminology. Amsterdam: John Benjamins.

Buitelaar P. and Sacaleanu B., (2001). Ranking and Selecting Synsets by Domain Relevance. In Proceedings NAACL WordNet Workshop.

Cabré, M. T., R. Estopà and J. Vivaldi (2001). Automatic Term Detection: A Review Of Current Systems. In Bourigault, D. et al. (eds) Recent Advances in Computational Terminology. Chp 3. Amsterdam: John Benjamins

Estopà, R. (1999). Extracció de terminologia: elements per a la construcció d’un SEACUSE (Sistema d’extracció automàtica de candidats a unitats de significació especialitzada). PhD thesis. Universitat Pompeu Fabra.

Kageura, K. and B. Umino (1996) Methods of automatic term recognition: A review. Terminology. 3:2. (pp 259-289).

Magnini B. and G. Cavaglià, (2000). Integrating Subject Field Codes In WordNet. In Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Greece.

Missikoff , M. Navigli, R. Velardi, P. (2002) The Usable Ontology: An Environment for Building and Assessing a domain ontology. Proceedings of International Semantic Web Conference, Sardinia, Italy, June 2002.

Montoyo, A. Palomar and M. Rigau, G. (2001). WordNet Enrichment with Classification Systems. In Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations. Pittsburg, 2001.

Vivaldi, J., (2001). Extrakció de Candidatos a Témino mediante combinación de estrategias heterogéneas. PhD thesis. Universitat Politècnica de Catalunya.

Vivaldi, J. and Rodriguez, H, (2002). Medical Term Extraction using the EWN ontology. In Proceedings of Terminology and Knowledge Engineering (pp 137-142). Nancy.