Clustering of Polysemic Words

Laurent Cicurel[1], Stephan Bloehdorn[2], and Philipp Cimiano[2]

[1] iSOCO S.A., ES-28006 Madrid, Spain
lcicurel@isoco.com
[2] Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany
{bloehdorn,cimiano}@aifb.uni-karlsruhe.de

Abstract. In this paper, we propose an approach for constructing clusters of related terms that may be used for deriving formal conceptual structures in a later stage. In contrast to previous approaches in this direction, we explicitly take into account the fact that words can have different, possibly even unrelated, meanings. To account for such ambiguities in word meaning, we consider two alternative soft clustering techniques, namely Overlapping Pole-Based Clustering (PoBOC) and Clustering by Committees (CBC). These soft clustering algorithms are used to detect different contexts of the clustered words, resulting in possibly more than one cluster membership per word. We report on initial experiments conducted on textual data from the tourism domain.

1 Introduction

Since many years, conceptual structures have played a major role in the construction of knowledge management applications. Instantiations of these include highly formalized ontologies, less formal taxonomic structures and, even less formal, groups of descriptors having intuitively similar interpretations. While ontologies (cf. Staab and Studer(2003)) show their full potential in knowledge based systems and reasoning engines, taxonomic structures have been mostly applied as structured controlled vocabularies in the context of library science as well as background knowledge in information retrieval, text mining and natural language processing (see e.g. Bloehdorn et al. (2005)). Plain term clusters have recently attracted attention as a means for structuring descriptors in social tagging systems.

Though all these conceptual structures can provide potential benefits for an increasing number of applications, their construction requires a costly modeling activity, a problem typically referred to as the *knowledge acquisition bottleneck*. Recent work in *ontology learning* (cf. Mädche and Staab (2001),

Acknowledgements: This work was supported by the European Commission under contract IST-2003-506826 SEKT and the by the German Federal Ministry of Education, Science, Research and Technology in the project SmartWeb.

Buitelaar et al. (2005)) has started to address this problem by developing methods for the automatic construction of conceptual structures. This is typically done in an unsupervised manner on the basis of text corpora relevant for the domain of interest. A major focus of these approaches has been the usage of term clustering techniques (see e.g. Grefenstette (1994), Faure and Nedellec (1998), Gamallo et al. (2005) and Cimiano et al. (2005)). However, a common weak point of these approaches is that they rarely take into account that words are *ambiguous*, i.e. they can have several – possibly grossly unrelated – meanings. Thus, in most approaches the assignment of words to clusters is assumed to be functional. An exception to this is certainly the work of Pantel and Lin (2003), which also provides the basis for our investigations in the sense that we use *soft clustering algorithms* which can assign words to different clusters, therefore accounting for their different contextual meanings. We restrict our attention on a flat clustering of terms, i.e. we do not aim at constructing a hierarchical structure between the term clusters - the work reported in this paper is thus meant to be a first step in ontology learning. Our contribution lies in the analysis of two different algorithm with respect to their ability to account for several meanings of words.

The remainder of this paper is structured as follows. After a quick review of the Distributional Hypothesis, Section 2 describes the feature representation of terms employed in our approach. In Section 3, we describe the two soft clustering algorithms used in our experiments, namely *Overlapping Pole-Based Clustering (PoBOC)* and *Clustering by Committees (CBC)*. In Section 4, we outline an approach for evaluating clusters of (ambiguous) words with membership in multiple clusters using WordNet as the corresponding gold standard. In Section 5, we report on results of an initial evaluation experiment on a tourism-related corpus consisting of texts obtainedfrom the 'Lonely Planet' website. We conclude in Section 6.

2 Term Representation

In this section, we give a short overview of the representation of terms in vector space of syntactic dependencies that will be used to apply the soft clustering techniques described in the next section. Hereby, we adopt the approach described in Cimiano et al. (2005), which is motivated by the *Distributional Hypothesis* (Harris (1968)). The Distributional Hypothesis claims that terms are semantically similar to the extent to which they share similar syntactic contexts. This means that, if two words occur in similar contexts, they are assumed to have a similar meaning. A syntactic context could be, for example, a verb for which the term in question appears as subject or object.

For this purpose, for each term in question, we extract syntactic surface dependencies from the corpus. These surface dependencies are extracted by matching texts tagged with part–of–speech information against a library of patterns encoded as regular expressions. Note that our approach is related to

the Generalized Vector Space Model (Wong et al. (1985)) but uses syntactic features instead of plain occurrences of words in documents. In our approach, first the corpus is part-of-speech tagged[3]. The part-of-speech tagger assigns the appropriate syntactic category to every token in the text. Features are then extracted by matching regular expressions defined over tokens and part-of-speech tags which denote syntactic dependencies between a verb and its subject, an adjective and the modified noun and the like. In what follows, we list the syntactic expressions we use and give examples of object–attribute pairs extracted. We employ predicate notation $a(o)$, where a is the attribute and o the object:

- adjective modifiers: e.g. *a nice city* → nice(city)
- prepositional phrase modifiers: e.g. *a city near the river* → near–river(city) and city–near(river), respectively
- possessive modifiers: e.g. *the city's center* → has–center(city)
- noun phrases in subject or object position: e.g. *the city offers an exciting nightlife* → offer–subj (city) and offer–obj(nightlife)
- prepositional phrases following a verb: e.g. *the river flows through the city* → flows–through(city)
- copula constructs: e.g. *a flamingo is a bird* → is–bird(flamingo)
- verb phrases with the verb *to have*: e.g. *every country has a capital* → has–capital(country)

The plain feature extraction gives each feature the same importance. However, it is certainly the case that not all features for a noun are equally representative. Thus, we replace the simple appearance count of a word with a feature by their pointwise mutual information value (cf. Church and Hanks (1990)). The feature vector space is high-dimensional and sparse. To increase the statistical properties, we have thus pruned features and words in our experiments, i.e. we considerd only those words which have at least a given number of features and the features that describe at least a given number of words.

3 Soft Clustering Algorithms

Clustering words which are potentially ambiguous into semantically homogeneous groups requires two main properties: (i) the clustering algorithm must allow clusters to overlap, i.e. a word can belong to one or more clusters and (ii) it needs to automatically determine an appropriate number of clusters. In our experiments, we have employed two soft-clustering algorithms, namely CBC, an algorithm which was developed by Pantel and Lin (2003), and PoBOC, developed by Cleziou et al. (2004). Note that we define as a *soft-clustering*

[3] In our experiments, we have used TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`)

algorithm every clustering algorithm for which the assignment of objects to clusters is non-functional. For soft-clustering algorithms the clusters thus typically overlap.

Overlapping Pole-Based Clustering: PoBOC

PoBOC is an overlapping cluster algorithm developed by Cleuziou et al. (2004). The similarity measure between two words we have used is the cosine of the angle between their vectors. The main idea of PoBOC is to find so called *poles* at a first step. Poles are very homogeneous clusters which are as far as possible from each other. The elements in the poles can be seen as *monosemous* words, i.e. words which have only one sense or meaning. After this first phase, several words remain unassigned though. These unassigned words are words which, as they do not form part of any pole, potentially feature several meanings and thus could end up in several clusters. After this pole construction phase, the remaining words are assigned to one or many poles.

Clustering by Committee: CBC

CBC was developed by Pantel and Lin (2003). CBC shares its two main phases with PoBOC: first it finds base clusters (poles in PoBOC and *committees* in CBC) and assigns the monosemous words to these committees. The committee construction is a recursive process which applies a hierarchical clustering algorithm on the k most similar neighbors in vector space of each word. Relying on an intra-cluster evaluation, only the best cluster is selected. The committees computed in this way are then filtered using the intra-cluster score and a threshold θ_1 that makes sure that the committees are far enough from each other. The process is then recursively applied to a *residue list* consisting of those words which are far enough from the committees with respect to a threshold θ_2 and thus can not be assigned to any committee.

In the assignment phase, every word is assigned to its most similar committee. Further, it is also assigned to the next most similar committee provided that the similarity is above some threshold θ_3 and that the similarity to the committees the word has been already assigned to is below some threshold θ_4. Hereby, the trick is that the non-zero features which are common to the word and the new committee to which it is assigned are set to '0' in the word vector, thus making sure that the word is always assigned to 'orthogonal' senses at later steps. Overall, CBC thus requires four parameters to be set, in contrast to PoBOC, which is parameter-free. As for PoBOC, we apply the cosine measure to assess the similarity between two terms.

4 Evaluation Methodology

In order to evaluate how well the produced clusters correspond to the actual different senses of the word, we compare the clusters with WordNet (cf. Miller

(1995)) as a *gold standard*. WordNet is a lexical ontology organized in interconnected synonymous groups of terms called *synsets*. The intuition behind our evaluation methodology is to compare the derived clusterings to the set of synsets defined in WordNet. Obviously, in the ideal case, the produced clusters would be one-by-one copies of the synsets. The procedure to evaluate the term clusters is as follows: first we assign each cluster to one or more synsets (depending on which of the two evaluation modes described below we consider) and then we approximate the semantic similarity between clusters and synsets by comparing the overlap between the words in the cluster and the words in the synsets using a vector-space model[4]. This is approach described in more detail below.

First method: One-To-One Association

This first method consists of assigning each cluster to exactly one synset, i.e. to the most similar one. Hereby the similarity is calculated as follows: for both the cluster and the synset, binary vectors are constructed which are then compared relying on the cosine similarity. The dimension of these vectors corresponds to the union of the words appearing in the cluster or the synset and the value of a dimension of the word/synset vector is 0 in case the corresponding word does not occur in the word/synset, 1 if it does occur. The score of a cluster is then calculated as its similarity with respect to the synset it has been assigned to. Intuitively, a high score means that the cluster resembles very much a synset that is actually defined in WordNet while a low score indicates that even the most similar synset achieves only a small similarity value indicating that the cluster comprises many different synsets. We define the score of the overall clustering as the average of the individual cluster scores.

Second method: One-to-several Association

The previous approach has the advantage of being simple and efficient to compute, but neglects the fact that WordNet is organized hierarchically. According to the above evaluation method, a clustering with clusters being composed of several synsets might still be considered a good one as long as the contributing synsets are semantically close within WordNet. The second evaluation method is thus similar in principle to the first one but takes into account the above intuition by the fact that clusters can be assigned to one or more synsets. Consequently, synset vectors are built possibly taking into account the words of more than one synset. As above, the overall score is the average score of all

[4] It is important to mention that this allows to assess the 'precision' of our clustering, but not the 'recall' , i.e. how many of the actual senses of a word we actually are able to account for. In any case, an evaluation in terms of recall-inspired metrics is quite problematic as not all the senses of a given word contained in WordNet are relevant for all domains.

clusters. The procedure to accomplish the assignment of a cluster to several synsets is as follows: first the cluster is assigned to its most similar synset as explained above. Further, it is iteratively also assigned to the next most similar synset provided that the score does not decrease and that the new synset is not to dissimilar from the original synset. Hereby, similarity between synsets is calculated relying on similarity measure introduced by Lin (1998), which is defined as:

$$sim_{lin} = \frac{2 * IC(lcs(syn_1, syn_2))}{IC(syn_1) + IC(syn_2)}.$$

Hereby *lcs* denotes the least common subsumer, i.e. the most specific common hypernym of the compared synsets syn_1 and syn_2 and $IC(syn)$ denotes the *Information Content* of a synset given by $IC(syn) = -\log(P(syn))$, where $P(syn)$ is the probability of encountering the synset estimated based on corpus frequency counts.

5 Experimental Results

As corpus for our experiments, we use a collection of approximately 1,000 texts downloaded from the LonelyPlanet website describing tourist destinations. The corpus is thus small, consisting of around 523,780 tokens. From this corpus, we extract 10,935 nouns with 19,218 features. Restricting on the nouns that have at least two features and the features that describe at least two nouns, we have an input to the clustering algorithm of 3,769 nouns with 5,041 features. The number of clusters as well as the average cluster size produced by each algorithm, PoBOC and CBC with the parameters $k = 10$, $\theta_1 = 0.35$, $\theta_2 = 0.30$, $\theta_3 = 0.01$, and $\theta_4 = 0.2$, is summarized in Table 1. Further, these numbers are compared to the average synset size in WordNet. We observe that PoBOC outputs bigger (and therefore less) clusters; the standard derivation of its clusters is also higher.

	No. clusters	Avg (stddev) cluster/synset size
PoBOC	1162	2.90(±2.10)
CBC	2010	1.68(±0.87)
WordNet	7897	2.14(±1.55)

Table 1. Basic Clustering Statistics.

Since this study is mainly concerned ambiguous words, it is interesting to get an idea of the average number of meanings of the clustered words. Considering the number of synsets a word belongs to as the number of meanings, a word has in average 3.42 meanings. In the PoBOC clusters, 92.9% of the terms have more than one meaning, and only 84.5% in the CBC cluster set. In

average, 71.4% of the terms in a cluster are ambiguous in the PoBOC cluster set, whereas 70.1% are amiguous in the CBC cluster set. It is also possible to count the number of clusters a word belongs to and use this value as the number of meanings of this word. Surprisingly, PoBOC and CBC obtain the same average score of 1.25 meanings, though they do not arrange the words the same way and do not detect the same words as ambiguous.

	Average cluster similarity	
	One-to-one	One-to-Several (0.5)
PoBOC	0.645	0.648
Not associated	1 cluster	1 cluster/6716 synsets
CBC	0.750	0.752
Not associated	14 clusters	14 clusters/5879 synsets

Table 2. Average evaluation scores and number of non-associated clusters.

The results of our evaluation in terms of average similarity as described in the previous section are presented in Table 2 for both evaluation modes. In the *One-To-Several*-association mode, we set 0.5 as similarity threshold which needs to be exceeded in order for a synset to be added to the evaluation set. In both modes, CBC obtains higher scores than PoBOC. There are at least two reasons to explain these results. First, CBC has a special mechanism designed to cluster words, namely the suppression of the word features in common with the committee, which avoids that words are assigned to distinct but very similar committees. The second reason is that CBC typically creates far more (2,010 versus 1,162) and consequently smaller clusters (average size of 1.68 versus 2.90) than PoBOC resulting in a slight bias in favor of CBC. This bias is due to the fact that the average size of the clusters produced by CBC is closer to the average synset size in WordNet.

We can thus conclude that the clusters produced by CBC correspond better to senses contained in WordNet than those produced by PoBOC. The disadvantage of CBC is certainly that a considerable number of parameters needs to be fixed or tuned, thus making its usage not as straightforward compared to PoBoC.

6 Conclusion and Future Work

We have analyzed two soft-clustering algorithms, CBC and PoBOC, with respect to the task of capturing the different corpus-specific senses or meanings of a word based on the Distributional Hypothesis and a corresponding representation of terms as vectors of syntactic features. We have further proposed an approach for the evaluation of this and related approaches and reported on experimental results for a dataset for the tourism domain. Our results indicate that clustering using CBC is more adapted for our purposes although we

pointed out that PoBOC has the advantage of having little parameterization, which makes it easier to use.

In future work, we aim at using alternative soft clustering approaches, possibly combining CBC and PoBOC. As both algorithms share the same phases, this seems definitely reasonable. Further, it also seems promising to examine a bootstrapping approach in which words are first assigned to clusters corresponding to their different meanings and then the different contexts provided by the clusters are used for disambiguation, yielding sense-specific features.

References

BLOEHDORN, S., CIMIANO, P. and HOTHO A. (2006): Learning Ontologies to Improve Text Clustering and Classification *Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005)*. Springer.

BUITELAAR, P., CIMIANO, P. and MAGNINI, B. (eds.) (2005): Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press.

CHURCH, K., and HANKS, P. (1990): Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics 16:22-29*.

CLEUZIOU, G., MARTIN L., VRAIN, C. (2004): PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04)*.

CIMIANO, P., HOTHO, A. and STAAB, S. (2005): Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research (JAIR), 24, 305–339*.

FAURE, D. and NEDELLEC, C. (1998): A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology. *Proceedings of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*.

GAMALLO, P., AGUSTINI, A. and LOPES, G.P. (2005): Clustering Syntactic Positions with Similar Semantic Requirements. *Computational Linguistics, 21, 107–145*.

GREFENSTETTE, G. (1994): Explorations in Automatic Thesaurus Construction. Kluwer.

HARRIS, Z. (1968): Mathematical Structures of Language. Wiley.

LIN, D. (1998): Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*.

MÄDCHE A. and STAAB S. (2001): Ontology Learning for the Semantic Web. *IEEE Intelligent Systems, 16, 72–79*.

MILLER, G.A. (1995): WordNet: a Lexical Database for English *Communications of the ACM, 38, 39–41*.

PANTEL, P. and LIN, D. (2002): Discovering Word Senses from Text. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*.

STAAB, S. and STUDER, R. (eds.) (2003): Handbook on Ontologies. Springer.

WONG, S. K. M., ZIARKO, W. and WONG, Patrick C. N. (1985): Generalized Vector Spaces Model in Information Retrieval. *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1985)*.