A similarity measure for unsupervised semantic disambiguation

Roberto Basili, Marco Cammisa, Fabio Massimo Zanzotto

University of Rome Tor Vergata,
Department of Computer Science, Systems and Production, 00133 Roma (Italy),
{basili,cammisa,zanzotto}@info.uniroma2.it

Abstract

This paper presents an unsupervised method for the resolution of lexical ambiguity of nouns. The method relies on the topological structure of the noun taxonomy of *WordNet* where a notion of *semantic distance* is defined. An *unsupervised* semantic tagger, based on the above measure, is evaluated over an hand-annotated portion of the British National Corpus and compared with a supervised approach based on the *Maximum Entropy Model*.

1. Introduction

Semantic disambiguation is a critical task in most NLP applications and it has been often analysed under different perspectives. It clusters a variety of specific processes that range from word sense disambiguation to named entity recognition and classification. In all such lines of investigation lexical semantic resources have always played a central role as static source of information or as paradigms of lexical representation able to inspire the disambiguation model themselves. Recently a number of works have adopted Wordnet as target repository but, at the same time, have been using it as a source of wider information than sense: in (Abney and Light, 1999) semantic relationships in Wordnet (i.e. hyponimy) is modeled in a probabilistic setting and the traversing of the hierarchy is seen as a Markov process enabling a variety of statistical inferences about lexical preferences and disambiguation. On a similar light more recent works (Ciaramita et al., 2003) apply a different learning technique over the Wordnet hierarchy structure to complement sense descriptions with hyperonim information in order to increase the accuracy of word sense disambiguation. A common feature of these studies is the role of the lexical hierarchy as the main trigger of the decision function, that is the critical source evidence for disambiguation.

In this paper we propose a similarity measure aimed to support an unsupervised approach to semantic tagging. This proposal represents a variant of the notion of *Conceptual Density* previously suggested as a tool for sense disambiguation (Agirre and Rigau, 1995). However, the major difference is the learning framework in which this measure applied to the Wordnet hierarchy enables a "natural" corpus-driven empirical estimation of lexical and contextual probabilities for probabilistic semantic tagging.

The basic assumption is that similar syntactic behaviour of words is due to similiarity on a semantic ground. This hypothesis (in a Bayesian perspective) can be inverted and rules for disambiguation (i.e. $prob(sense|word)$) can be developed by generalizations over similar syntactic cases, i.e. several words in the same grammatical contexts. Such generalizations are useful semantic explanations for the underlying syntactic phenomena. We could refer this hypothesis as "one sense for syntactic collocation" in line with previous successful works (Yarowsky, 1995).

Finally, a last (but not least) principle is enforced. Within one collection (i.e. a domain) words tend to exhibit a reduced set of their own senses. A good general strategy for disambiguating one word is to find the minimal set of senses able to "explain" all its syntactic behaviours. In this perspective, first local contexts are analysed and suitable explanations are generated. Then only the best ones are preserved after most of the different contexts have been analysed. In a probabilistic setting, first probabilities of senses in specific syntactic configurations (i.e. $prob(sense|word, synt\_context)$) are computed. Then, from the different "local" scores, an overall (i.e. common to all the corpus) distribution is derived for the target word (i.e. $prob(sense|word)$). This is a probabilistic interpretation of a "one sense per domain" principle. Estimations at a "local" level thus exploit the conceptual density measure to favour some senses (i.e. explanations) over other. Then the global effect is to increase scores for the most effective explanations, those senses emerging from most of the local phenomena. The result is an unsupervised approach to disambiguation in line with previous results in this area.
Experimental results over an hand-annotated portion of the British National Corpus (about 5 M words) have been obtained (see a related abstract (Guthrie et al.2003) also submitted to LREC 2004). Although below the results obtained by a supervised method (ME approach to hand labelled data), the proposed unsupervised tagger confirms the effectiveness of the proposed notion of conceptual density as well as show a promising direction.

2. A semantic similarity measure for Unsupervised Tagging

The idea behind the suggested approach is that using Wordnet, a large Corpus C and a reference set of semantic categories $s \in SemList$, we can automatically estimate the lexical probability $p(s|tw)$ and use it for tagging. Notice that as the acquisition of these probabilities is dependent in general on the originating syntactic context r, $p(s|tw)$ are in fact derived from $p(s|tw, r)$. The idea is that **if** large evidence about syntactic phenomena r can be collected from the corpus **then**

- semantic similarity is constrained within the set of words in similar syntactic dependencies r and each dependency suggests a specific aspect of the word meaning

- syntactic dependencies are *independent*, so that global probability can be computed by summarizing over the entire set of relations r observed for the target words tw.

The above properties are exploited in the light of the semantic description offered by a reference lexical semantic resource, like Wordnet. Given a set of nouns W (determined by the syntactic collocation, r), a measure of the suitability of their generalizations t for r is the *information density* of the subtree in the hierarchy rooted at t. The higher is the density of the dominated tree t (i.e. the larger is the number of nodes in t acting as useful generalizations of some nouns in W), the better is the related generalization. Due to ambiguity of nouns, several trees can be built to cover different and overlapping subsets of W. In Fig. 1 two branches of the lexical hierarchy activated by four nouns $w_1, ..., w_4$ are depicted: the rightmost is a better representation of the nouns. It dominates in fact all nouns and most of its nodes are useful generalizations of more than one noun (e.g. node 9 that is a generalization of nouns w_1, w_3, w_4). The measure of density here employed will be herafter referred to as conceptual density.

We would like to automatically select the set of nodes like 9 that cover all the nouns by dominating a maximally dense subtree. One noun may well preserve more than one sense within different trees, whenever such different senses still apply as explanations of r. Given a syntactic collocation r and a corresponding set W, a greedy algorithm has been thus applied to generate the set of dominating trees t with the maximal conceptual density and able to "cover" every noun in W. Each noun is attached to a generalization through a conceptual density score. These scores are proportional to the system confidence into their underlying useful senses s and their probabilities, $p(s|tw, r)$. The scoring method for generalizations (i.e. hyperonims), as applied to Wordnet lexical hierarchy, is defined in the next section.

2.1. Word similarity among "syntactically equivalent" nouns.

Given a specific syntactic relation r, the likelihood of a sense s for tw is proportional to the number of other words in r that have common generalization with tw along paths activated by s in LKB. Each generalization will suggest a useful interpretation of r. As too much general interpretations are not useful (as they do not support effectively separation of word senses) the "most specific" among the "common" generalizations should be preferred. Notice that every choice (i.e. the selection of a subtree of the lexical hierarchy) represents a compromise between "*coverage*" (i.e. commonalities among different words) and "*specialization*" (i.e. granularity/specificity of the detected sense). A measure for capturing the quality of interpretations of nouns (i.e. their generalizations through the hierarchy) is the information density that a subtree provides with respect to the target set to be covered. The higher is the density of the selected tree (i.e. most of its nodes are useful as they are representing nouns in the target set) the better is the related generalization (i.e. the synset s root of the tree). Notice that such density has been used elsewhere (Agirre and Rigau, 1995) as a similarity measure for semantic disambiguation in sentences while we intend here use it as a score for ranking alternative sense of a word, given a syntactic context r.

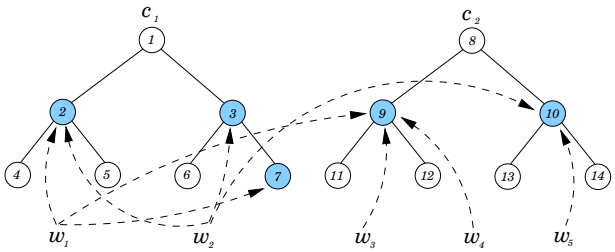

Figure 1: An example tree

DEF (*Conceptual Density*). Given: a syntactic collocation r, that determines a target set of nouns w, a synset s in Wordnet used to represent N lemmas in r, the conceptual density, $cd^{(r)}(s)$, of s with respect to r is defined by the following:

$$cd^{(r)}(s) = \frac{\sum_{i=0}^{h} \mu^i}{area(s)} \quad (1)$$

where :

- h is the estimation of the depth of a tree able to represent the N nouns. Its actual value is estimated by:

$$h = \begin{cases} \lfloor log_\mu N \rfloor & \textbf{iff } \mu \neq 1 \\ N & \textbf{otherwise} \end{cases} \quad (2)$$

- μ is the average number of sons per node in the actual Wordnet subhierarchy dominated by s. Its estimation is available statically from Wordnet and can be evaluated *a priori* without uncertainty. Notice that when nodes belong to unbalanced branches of the hierarchy, the value for μ can approach (and infact is) 1, so that a specific treatment of them is needed in the definition.

- $area(s)$ is the number of nodes in the s subhierarchy. This value is also estimated statically from Wordnet.

Notice that definition in Eq. (1) provides a notion of conceptual density slightly different from the earlier proposal of (Agirre and Rigau, 1995). In particular the number of covered nouns N is here used in the estimation of the tree in a rather different way that does not longer requires adjustment parameters as suggested in the original work. It should be also noticed that the original formula of CD :

$$cd^{(r)}(s) = \frac{\sum_{i=0}^{n} (\mu^i)^\alpha}{area(s)}$$

(where n is the number of nodes activated), brings to different results. If is considered the tree in Figure 1, the original formula of CD gives for the first tree the value 1, and for the second $\frac{3}{7}$, it can be induced that the set of these two trees must be used to cover all the five words. On the contrary, with the notion of CD defined in Eq. 1, the values will be respectively $\frac{1}{7}$ and $\frac{3}{7}$. It is clear, that with only the second tree (i.e. the tree with the higher value of CD) is sufficient to cover and then to explain all the words considered. A consequence is that the probability of the sense 9 for w_1 ($p(9|w_1)$) is higher than the others 2 ($p(2|w_1)$) and 7 ($p(7|w_1)$).

Now the Equation 1 applies to any valid and common gneralizations of the nouns in the target set. The aim of this method however is to reduce the number of such generalizations as much as possible. It has been thus defined as a useful set of generalizations for a target set r of Wordnet synsets $S = \{s \in LKB | s$ is an hyper. of at least 1 sense$s(w), w \in r\}$ such that:

$a)$ s is an hyperonim of at least two words in S

$b)$ the factor $\sum_{s \in S} cd_{(r)}(s)$ is maximal, among the different S in the family of sets S' that satisfy $a)$.

The optimal set O can be found by the following *greedy* algorithm.

1. Let the output set of synset O be the empty set, i.e. $O = \emptyset$.
2. Let S be the set of all synsets s that satisfy property $a)$ with respect to W.
3. Rank elements in S according to decreasing values of $cd^{(r)}(s)$.
4. While W and S are not empty
   (a) Let $s \in S$ be the highest ranked element
   (b) Let $C \subset W$ be the set of nouns whose senses are hyponims of s
   (c) $W = W - C$
   (d) $S = S - \{s\}$
   (e) If $C \neq \emptyset$ then $O = O \cup \{s\}$
5. Return(O).

The outcome of the above algorithm is the set O of synsets that are *the maximally dense generalizations of at least two words* in W. If a word w has no such generalization, it will not be represented in the resulting set O

The estimation of the probabilities in the algorithm for the tagging problem requires more than Equation (1). In fact the above algorithm and the Eq. 1 generate the set of senses O useful to interpret nouns $w \in W$ and assign them a score (confidence factor in their usefulness as potential generalizations of nouns wrt r). To map conceptual density into a probability, a maximum likelihood approach is applicable as follows.

DEF (*lexical probabilities*) Given a target set r and a word w, with its senses $s_1, ..., s_k$ in the lexicon, the conceptual density defined in Eq. (1) provides a score $cd^{(r)}(s)$ for the generalizations s of senses s_i of w. One or more senses, s_i, is interested by such a score if it is dominated by s in the hierachy. So s_i are individual senses of w while s are their generalizations. We accumulate the conceptual density of generalizations s over senses s_i dominated by them and then normalize. The following probability distribution can thus be defined:

$$prob(s_i|w, r) = \frac{\sum_{s \text{ hyper. of } s_i} cd^{(r)}(s)}{CD(w, r)} \qquad (3)$$

where

$$CD(w, r) = \sum_{j=1}^{k} \sum_{s \text{ hyper. of } s_j} cd^{(r)}(s)$$

3. Empirical Evidence

A way to evaluate the accuracy of the semantic generalization carried out by the algorithm of the previous section is to apply it to semantic disambiguation and tagging. The proposed similarity measure has been applied to tagging an extensive annotated portion of the BNC, where contexts like: $(tw\ r_1\ r_2\ ...\ r_k ...)$ have been studied to assign the semantic class that better explain the semantics of the target word tw. Statistical word tagging according to the estimated probabilities has been firstly applied and then compared against high general semantic categories (similar to those adopted in *LDOCE*). Either the LDOCE to Wordnet mapping and the tagging tasks have been accomplished in an automatic way thanks to the usage of the conceptual density measure. The unsupervised model has been also extended by back-off (Katz, 1987) in order to deal with unseen phenomena. Tagging results obtained over two different BNC data sets allowed a comparison between automatically assigned tag and two baselines (i.e. random choice and first WN sense). The accuracy of the proposed method (about 82%) is higher than the performance of the last baseline.

3.1. Semantic Tagging the British National Corpus

Most statistical models of NLP tasks (e.g. HMM in POS tagging) apply supervised approaches. This is also viable in *Semantic Tagging* (ST): available annotated texts can be used to derive general rules and apply them to new incoming texts. On the other side, ST making use of external lexical resources as static source of information (i.e. the sense, or class, dictionary), for inducing tagging preference rules are also possible. In these latter methods, resources reduce the need of large sets of annotated examples, i.e. they enable weakly supervised methods.

In the John Hopkins 2003 Summer Workshop "*Semantic Analysis Over Sparse Data*"[1], an unsupervised approach to tagging has been investigated based upon the similarity measure defined in the previous sections. The workshop gave us the possibility to rely on an human annotated corpus which contained 198,970 noun phrases, produced as a subset of the British National Corpus (Burnage and Dunlop, 1992). On the annotated head nouns the inter-annotator agreement was about 94%. About $13,097$ instances were set aside as a blind corpus ($Blind$). Experiments were performed which used the remainder of the human annotated corpus as training, and other experiments were unsupervised.

Unsupervised experiments used Wordnet as a basic resource for estimating source probabilities then a simple probabilistic model to annotate the test cases. Supervised approaches used Maximum Entropy methods primarily over annotated data extended with the results of parsed data (e.g. modifying adjectives and/or verbal heads), or with topical information (e.g. the topics of the source documents). The method employed for unsupervised tagging is based on grammatical parsing of the target corpus (i.e. extraction of basic dependencies involving the target head

[1] see URL at `http://www.clsp.jhu.edu/ws2003/groups/sparse/`

nouns), on measures of semantic similarity over the hierarchy and on a *back-off* tagging model. The training process can be summarized as follows:

1. Parse the training corpus and extract syntactic couples and triples whose head or modifier is the head noun target for semantic tagging. In this phase *word-preposition-noun* triples (e.g. *to_drink-during-dinner*, *water-with-gas*) or *verb-direct_object*, *subject-verb* couples (e.g. *drink-water*, *boy-drink*) are derived.
2. Then classes of nouns are derived by fixing the grammatical heads and syntactic relations r: for example, all the direct objects of the verb *drink* are clustered in the set $W^r = \{beer, water, ...\}$.
3. Assign preferences to senses l of the different nouns (e.g. $water$) within the specific grammatical context r (i.e. as objects of the verb $drink$), following the measure $cd^r(l) = \Sigma_{C \prec l} cd^r(C) \quad \forall tw \in W^r$ introduced in Eq. 1.
4. Estimate local probabilities $p(l|tw, r)$ out from the preferences $cd^r(l)$ derived at the previous step.
5. Estimate global probabilities $p(l|tw)$, $p(l|r)$ and $p(l)$

After training, tagging a noun tw within an incoming, grammatically analyzed, sentence:

$$tw \quad r_1 \quad r_2 \quad \cdots \quad r_k$$

is carried out by maximizing the probability $p(l|tw, r_1 \cdots r_k)$ and then map it into a target category C (i.e. the coarse grain category[2] used for annotating the BNC). The following formula shows how a simple probabilistic method based on *back-off* (Katz, 1987) has been applied to derive the most likely sense l of tw in the context $r_1, \ldots, r_k$: $p(l|tw, r_1, \cdots, r_k) = \prod_{i=1}^{k} p(l|tw, r_i)$. Where the back-off model is used when one or more probabilities are unknown or unreliable, in these cases a weaker probability is used. So to have for $p(l|tw, r)$ an approximation derived by a linear combination $p(l|tw, r) = \alpha p(l|tw) + \beta p(l|r)$ made respectively with the lexical and the syntactic probability. In the worst case, these probabilities are derived by the corpus probability $p(l)$. A more detailed description of the back-off model is reported in (Basili and Cammisa, 2004b).

After the best sense l for the target word tw is available the tagger maps it into LDOCE categories according to the method discussed in (Basili and Cammisa, 2004a). Results for supervised methods (based on ME trained with topical information plus adjectival modifiers) were around 85% for an held-out data set of about $99,000$ cases and 93,4% on the $13,097$ cases (blind corpus). Table 1 reports the results of the unsupervised tagger over the blind corpus and over the Held-out. The assumed baseline is the algorithm that tags the corpus according to the first Wordnet sense (i.e. the sense assumed by the Wordnet authors as the most common for n). The third row tells us the number of correct decisions when both the first two solutions are accepted.

[2] In the experiments target categories came from the Longman Dictionary of Contemporary English (Procter, 1978), like *Human*, *Abstraction*, *Animal* or *Collective Human*)

Table 1: Performance of the Unsupervised Tagger.

Tagging Algorithm	Blind	Held-Out
Pick the 1^{st} sense	68,74%	72,40%
Unsupervised Tagger (*argmax*)	81,05%	75,45%
Unsupervised Tagger (coverage of 1^{st} 2 senses)	95,17%	91,28%

The major outcome is that unsupervised methods, not making use of annotated examples, are below the accuracy of supervised techniques but they are viable as converging towards high levels of performance. It is to be noticed that no actual large scale experiment in sense disambiguation or acquisition of selectional restrictions for verb arguments has been shown to outperform the "Pick the 1^{st} Wordnet sense" baseline, while the unsupervised tagger is well above this heuristics. Further exploration should study combinations of the Wordnet-based approach with the annotated material. Weakly supervision can be obtained by seeding the process with a small number of annotated cases and then adding external evidence to bootstrap to larger scales.

4. References

Abney, S. and M. Light, 1999. Hiding a semantic hierarchy in a markov model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, ACL*.

Agirre, E. and G. Rigau, 1995. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of the First International Conference on Recent Advances in NLP. – Tzigov Chark, Bulgaria, September 1995*.

Basili, R. and M. Cammisa, 2004a. An automatic mapping between ldoce and wordnet. In *Forthcoming*.

Basili, R. and M. Cammisa, 2004b. Unsupervised semantic disambiguation. In *Forthcoming*.

Burnage, G. and D. Dunlop, 1992. Encoding the British National Corpus. In *Proceedings of the 13th International Conference on English Language Research on Computerised Corpora*.

Ciaramita, Massimiliano, Thomas Hofmann, and Mark Johnson, 2003. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *Proceedings of 18th International Joint Conference on Artificial Intelligence.*.

Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transaction on Acoustics, Speech, and Signal Processing*.

Procter, P., 1978. *Longman Dictionary of Contemporary English*. Longman Group.

Yarowsky, David, 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Meeting of the Association for Computational Linguistics*.