AUTOMATIC DETECTION OF SEMANTICALLY PRIMITIVE WORDS USING THEIR REACHABILITY IN AN EXPLANATORY DICTIONARY *

GRIGORI SIDOROV and **ALEXANDER GELBUKH**

Natural Language Laboratory, Center for Computing Research (CIC),
National Polytechnic Institute (IPN), 07738, D.F., Mexico.
{gelbukh, sidorov}@cic.ipn.mx

Abstract We suggest the method that permits building a set of candidates to be considered semantic primitives from the standard explanatory dictionary. Our method is based on the frequencies of the words that are reachable in a semantic network constructed from the dictionary. The method implements word sense disambiguation techniques, network construction, and reachability analysis. In part of word sense disambiguation we use an improved Lesk's algorithm. In the part of analysis of reachability we show that the words to which our algorithm assigns high weight, are plausible candidates to be semantic primitives. It is also shown that better candidates to semantic primitives should be included in short vicious cycles, which is detected by our algorithm. We applied the method to a rather large Spanish explanatory dictionary.

1 Introduction

One of the problems of modern computational linguistics is the problem of defining vocabulary. Several types of dictionaries of English, for example, by Oxford or Longman, are using a restricted set of defining words. Usually the number of words in such set is about 2000 to 3000.

This practice is directly connected with the problem of semantic primitives that has a long history in linguistics. The history of the problem is very well known, so we do not discuss it here in detail. It is worth noting that on the modern stage the most consistent scientist is Anna Wierzbicka. The other approach to this problem has been developed by Yu. Apresjan.

Usually the normal words from a language are taken as primitives, though it is stressed that they are no more polysemic. Still, this point is not so clear because the polysemy is the core of the word sense. Thus, in theoretical aspect it cannot be left apart so easily. Nevertheless for the practical purposes we can assume that the primitives have only one meaning.

On the other hand, the ideas of regularity in meaning of the words are on the agenda. We think that the increasing interest to this theme is connected with the work (Pustejovski, 1991). Among the last works, we would like to mention [Ravin and Leacock, 2000]. Regular polysemy in WordNet that is represented by inheritance.

An explanatory dictionary defines words through definitions composed of other words, e.g., *bank is a financial institute*. This looks like a relation between the defining words and the words being defined. But it is not: in fact, what is defined are not words but word senses: *bank*$_1$ *is a financial institute*, while *bank*$_2$ *is the edge of a river*. However, the defining words are (in existing dictionaries) still strings rather than senses: in the definition of *bank*$_1$, is *institute* a school, a research center, a social structure, or an organization? Any NLP application of an explanatory dictionary requires sense disambiguation (WSD) in such definitions.

The problem of WSD is well investigated. The prevailing approaches are knowledge-poor statistical approaches (Manning and Shutze, 1999) based on bayesian classifiers, neural networks, support vector machines, and other purely statistical techniques.

On the other hand, knowledge-rich approaches were suggested as early as in (Lesk, 1986) and (Hirst, 1987). An advantage of knowledge-rich approaches is their clarity and explicitness: it is easy to see why the algorithm makes a decision and on what information the decision is based. Additionally, as more lexical resources become available, knowledge-rich approaches become more affordable. Because of this, some recent works have presented modifications of the original Lesk's algorithm based on the use of thesaurus, synonym dictionaries, different kinds of morphological normalization, etc. (Wilks and Stevenson, 1998, 1999), (Mahesh *et al*., 1997), (Cowie *et al*., 1992), (Yarowsky, 1992), (Pook and Catlett, 1988).

In Lesk's algorithm, a word sense is represented as the set of strings that form the definition of the sense: e.g., *bank*$_1$ = {"*financial*", "*institute*"}. The algorithm calculates the scores of a word sense on the basis of the intersection of this set with the senses of the words in the context and chooses the sense with the best scores. We improved the algorithm by introducing fuzzy comparison between such strings

* Work partially supported by CONACyT, SNI, and CGEPI-IPN.

Grigori Sidorov, Alexander Gelbukh. *Automatic Detection of Semantically Primitive Words Using Their Reachability in an Explanatory Dictionary*. IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE 2001, IEEE, pp. 1683–1687; www.Gelbukh.com/CV.

based on the use of a large synonym dictionary and derivational morphological normalization.

We apply this improved algorithm to a rather large Spanish explanatory dictionary (30,000 entries). This dictionary is not as "good" dictionary with restricted defining vocabulary as, for example, Longman Dictionary of Contemporal English.

Note that WSD in dictionary definitions is greatly simplified as compared with disambiguation applied to a usual text because in this case (1) tagging (which is usually the first step in WSD) is simplified since definitions are structured texts; also, the information on the grammatical category of the headword helps tagging; (2) all words in a definition are known to be related with the headword just because they are parts of its definition; and (3) the problem of context window size is not relevant since the whole definition is used.

In the rest of the paper, we will first describe the algorithm, then discuss the obtained results, and finally draw some conclusions.

2 Basic ideas of the method

Let us take as input a usual explanatory dictionary that does not implement any semantic improvement. The dictionary is processed in a manner that it is converted to a semantic network, i.e., every meaningful word is connected with the meaningful words that form its definition. Note, that we should disambiguate word senses in the definitions of words to be able to construct such a network. Now it is possible to investigate the relation of reachability between words in determined number of steps. We call the words N-reachable if they can be reached in such a network in a number of steps equal to N. The N-reachable words with maximum frequencies are candidates to be semantic primitives.

The other idea that can help in automatic extraction of the semantic primitives is the idea of vicious circles. Namely, the word should be the primitive if its definition has a vicious circle in the semantic network. By vicious circle we mean that there exist a root in the network that passes more than once through the same node. In fact, that means that there is no possibility to define this word using other words from the same dictionary.

3 The algorithm

The method has as an input a machine-readable dictionary. The first steps are traditional processing, such as tagging and word sense disambiguation (WSD). For tagging we used a method based on syntactic heuristics. For WSD we developed a method based on the Lesk's algorithm. We modified it by adding the model of words derivation and use of a synonym dictionary. The detailed description is given below.

The next step is construction of a semantic network and analysis of reachability. In our case we chose N = 3. Then the frequencies of 3-reachable words are counted.

Finally, we show that the method gives better results than the method without construction of a semantic network that is a baseline for our method. The latter method simply uses the frequencies of words in a dictionary. We prove that our method is better by comparing the number of words with vicious circles received in both methods. Our method gives more words with vicious circles (87% for the first 1000 words) as compared with the baseline (70% for the first 1000 words).

Now let us give a detailed description of the steps mentioned above.

3.1 Word sense disambiguation

For each word (string) in each definition, we look up this word in the same dictionary. If for this word there are several senses in the dictionary, the problem consists in the choice of the most plausible one. Our algorithm for the solution of this problem consists in two stages: preprocessing and scoring. Then, for each word, the sense with best scores is chosen.

Preprocessing. This stage consists of tagging (determining the part of speech of each word) and normalization (reducing of the word to a standard form).

Since the results of the tagger are the input of the next algorithm, any existing tagger can be used at this stage.

We used the tagger that was available for Spanish. For tagging, we use a set of syntactic heuristics developed specifically for the dictionary that we used (a Spanish dictionary). Some of the heuristics deal with the syntactic structure of a sentence, for example: a word preceded by an article (other than *el*) cannot be a verb. Another type of heuristics uses knowledge of the definition structure, for example: in the definition of a noun, the first word is a noun.

For normalization, we use a morphological system that reduces the words to a standard form, preserving its part of speech (like *teaches*, *taught*, *teaching* → *(to) teach*).

Also at this stage the functional words (stopwords) are excluded from the definitions.

Scoring. We represent each word sense as the set of words (without stopwords) that form its definition.

Let for a word (string) w in a definition of some sense (represented as a set h), several senses (represented as sets s_1, ..., s_n) are found in the dictionary. As the score of each sense s_i, we use the proximity measure between s_i with h defined as follows. Let a, b be two sets of words (strings), then the proximity measure $w(a, b) = \sum_{x \in a,\ y \in b} w(x, y)$, where $w(x, y)$ is the proximity measure between two words defined as follows.

If $x = y$, then $w(x, y) = 1$. Otherwise, if x is a synonym of y or y is a synonym of x, then $w(x, y) = 0.5$. Otherwise, if the initial parts (at least 5 letters long) of the two words coincide (e.g., x = *presidente* and y = *presidir*), we consider such words derivates of each other. The latter represents a very simple model of derivational morphology, which of course can be improved in the future. In this case also $w(x, y) = 0.5$. Otherwise, $w(x, y) = 0$.

3.2 Semantic network and statistics

Basically, our algorithm works as follows.

- Each dictionary entry is represented as a pair: the headword and a set of defining words.
- Our algorithm iteratively substitutes each defining word with its definition.
- After a given number of iterations, we count the number of occurrences of each word in the definitions of the resulting dictionary.

This algorithm tends to eliminate from the definitions the words that are defined through other words, i.e., no-primitive words. On the other hand, since the primitive words are necessarily parts of vicious cycles, their frequency tends to increase. Thus, the number of occurrences in the resulting dictionary after some number of iterations can be used as the measure of the "primitiveness" of the word.

In the full paper, the algorithm is explained in more detail.

4 Experimental results

We applied our method to a Spanish dictionary of about 30,000 entries, with the average number of words (without stopwords) per definitions being about 8.

4.1 Tagging and Word Sense Disambiguation

As a baseline, we also implemented and applied to the same dictionary two other algorithms: (1) the original Lesk's algorithm (without any fuzzy comparison, i.e., with $w(x, y) = 1.0$ when $x = y$ and $w(x, y) = 0$ otherwise) and (2) an algorithm that always chooses the first sense of the word.

Then we randomly chose 50 headwords and manually verified the results for their definitions. Our algorithm disambiguated incorrectly 13% of correctly tagged ambiguous words (i.e., without counting unambiguous words and words incorrectly tagged by the tagger). This is one-fourth better than the original Lesk's algorithm, which produced 17% of errors, and twice better than that for the algorithm that always chooses the first sense, which produced 29% of errors. (With counting also unambiguous words, these figures were 12%, 16%, and 28% of errors, correspondingly.)

In our test program, 92% of words were correctly tagged, the majority of errors consisting in confusing nouns and adjectives; clearly, for incorrectly tagged words correct disambiguation was impossible, this is why we did not count them in the statistics above. Incorrect part of speech tagging of a word did not affect much the disambiguation results for the other words in the same definition because of our morphologically-based comparison and because usually (in 75% of cases in our experiments) there is a little difference in the definition of a noun and the corresponding adjective.

4.2 Semantic network and statistics

We analyzed the words with reachability N = 3. The frequencies of these 3-reachable words were counted. The first 30 words are given in Appendix 1.

We compared the list of the 3-reachable words with the list of the words obtained directly from the dictionary, it is obvious that it is a list of 1-reachable words. This list serves as a baseline for our method.

To compare the results, we used the idea that the primitives should have vicious circles. The additional fact that testifies that this idea is correct is that in the lists of 3-reachable words the number of words with vicious circles is the least for the first thousand of the words, and increases with each next thousand words while in the 1-reachable list the number of these words decreases.

The number of words with vicious circles is given in the following table:

Frequency rank	Ordered by frequency of words	Ordered by frequencies of 3-reachable words
0001–1000	70%	88%
1001–2000	74%	70%
2001–3000	85%	65%

In Appendix 2, we list 30 words that do not have the vicious circles (note that it is so for the reachability N = 3, maybe for greater values the circles will appear).

5 Conclusions

We have suggested the method that permits building a set of candidates to be considered semantic primitives from the standard explanatory dictionary. Our method is based on the frequencies of the words that are reachable in a semantic network constructed from the dictionary. The method implements word sense disambiguation techniques, network construction, and reachability analysis. We applied the method to a large Spanish explanatory dictionary.

In part of word sense disambiguation we used an improved Lesk's algorithm. Our improvements consist in fuzzy comparison of words using a synonym dictionary and a simple derivational morphology procedure. Our algorithm gives better results than the original Lesk's algorithm and a baseline WSD algorithm. Also, application of disambiguation to dictionary definitions (as compared with usual texts) allows for considerable simplification of the algorithm.

In the part of analysis of reachability we show that the words to which our algorithm assigns high weight, are plausible candidates to be semantic primitives. We have also shown that better candidates to semantic primitives should be included in short vicious cycles, which is detected by our algorithm.

Appendix 1. The frequency list of the first 3-reachable words.

In this list we give the first 30 words that are 3-reachable. The original Spanish words are given in parenthesis. The English words are translated from Spanish, so sometimes we have several variants, like *to make/to do*.

703639 thing (cosa)
404368 person (persona)
338330 something (algo)
231366 to make/to do (hacer)
216705 set/group (conjunto)
187979 action (acción)
177864 to have (tener)
151232 body (cuerpo)
147475 part (parte)
129787 not (no)
96527 one (uno)
94933 form (forma)
87357 name (nombre)
85241 to give (dar)
81072 element (elemento)
77019 relative (relativo)
70406 word (palabra)
69836 object (objeto)
69484 effect (efecto)
68227 to form (formar)
66897 time (tiempo)
65257 group (grupo)
64761 grammar (gramática)
64207 to be able to (poder)
62690 certain (determinado)
61145 animal (animal)
60902 two (dos)
56288 to take place (producir)
47881 class (clase)
47708 human (humano)

Appendix 2. The list of 3-reachable words without vicious circles.

Here we give 30 samples of 3-reachable words from the first thousand that do not have vicious circles.

constitute (constituir)
muy (very)
dentro (inside)
entity (entidad)
chemical (químico)
majority (mayoría)
affirm (constar)
union (unión)
before (ante)
great (gran)
vegetable (vegetal)
electric (eléctrico)
limited (limitado)
corresponding (correspondiente)
bad (mal)
site (sitio)
study (estudiar)
ordered (ordenado)
diverse (diverso)
organization (organización)
celular (celular)
colorless (incoloro)
illness (enfermedad)
cavity (cavidad)
also (también)
situated (situado)
evoke (evocar)
cape (capa)
always (siempre)
correct (correcto)
component (componente)

References

Apresjan, J. D. (1974) Regular polysemy, Linguistics, 142: 5-32.

Apresjan, J. D. (1995) Selected works (in Russian). Moscow, V 1, 472 p., V 2, 768 p.

Budanitsky, A. (1999) *Lexical semantics relatedness and its application in natural language processing*. Technical report CSRG-390, University of Toronto, Toronto, 146 p.

Cowie J., Guthrie, L., and Guthrie, G. (1992) Lexical disambiguation using semantic annealing. Proceedings of *Coling-92*, Nante, France, pp. 359-365.

Evens, M. N. (ed.) (1988) *Relational models of lexicon: Representing knowledge in semantic network*. Cambridge: Cambridge University Press.

Fellbaum, C. (1990) The English verb lexicon as a semantic net. *International Journal of Lexicography* 3: 278-301.

Hirst, G. (1987) *Semantic interpretation and resolution of ambiguity*. Cambridge, Cambridge University Press.

Karov, Ya. and Edelman, Sh. (1998), Similarity-based word-sense disambiguation. *Computational linguistics*, Vol. 24, pp. 41-59.

Kozima, H. and Furugori, T. (1993) Similarity between words computed by spreading activation on an English dictionary. In: Proceedings of the 6 conference of the European chapter of ACL, pp. 232-239.

Kozima, H. and Ito, A. (1997) Context-sensitive word distance by adaptive scaling of a semantic space. In: Mitkov, R. and Nicolov, N. (eds.) Recent Advances in NAtural Language Proceedings: Selected Papers from RANLP'95, pp. 111-124.

Lesk, M. (1986), Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of *ACM SIGDOC Conference*. Toronto, Canada, pp. 24-26.

Mahesh, K., Nirenburg, S., Beale, S., Raskin, V., and Onyshkevich, B. (1997) Word sense disambiguation: Why have statistics when we have these numbers? Proceedings of *7th International Conference on Theoretical and methodological issues in machine translation*. Santa Fe, NM, pp. 151-159.

Manning, C. D. and Shutze, H. (1999), *Foundations of statistical natural language processing*. Cambridge, MA, The MIT press, 680 p.

McRoy, S. (1992) Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, Vol. 18(1), pp. 1-30.

Pook, S. L. and Catlett, J. (1988) Making sense out of searching. *Information outline 88*, Sydney, pp 148-157

Ravin, Y. and Leacock, C. (eds.) (2000) *Polysemy: Theoretical and computational approaches*. Oxford: Oxford University Press, 227 p.

Saint-Dizier, P. and Viegas, E. (eds.) (1995) *Computational lexical semantics*. Cambridge: Cambridge University Press, 447 p.

Wilks, Y. and Stevenson, M. (1998), Word sense disambiguation using optimized combination of knowledge sources. Proceedings of *ACL 36/Coling 17*, 1398-1402.

Wilks, Y. and Stevenson, M. (1999), Combining weak knowledge sources for sense disambiguation. Proceedings of *IJCAI-99*, 884-889.

WordNet: an electronic lexical database. (1998), C. Fellbaum (ed.), MIT, 423 p.

Wierzbicka, A. (1980) *Lingua Mentalis: The semantics of natural language*. New York: Academic Press.

Wierzbicka, A. (1996) *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Yarowksy, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceeding of *Coling-92*, Nante, France, pp. 454-460.