Word Sense Disambiguation in a Spanish Explanatory Dictionary

Grigori Sidorov and Alexander Gelbukh

Natural Language Laboratory,
Center for Computing Research,
National Polytechnic Institute
Av. Juan de Dios Bátiz, s/n, esq. Mendizabal,
Zacatenco, C.P. 07738, Mexico D.F., Mexico.
{gelbukh, sidorov}@cic.ipn.mx

Résumé – Abstract

Nous appliquons la désambiguïsation du sens des mot aux définitions d'un dictionnaire explicatif espagnol. Pour calculer le grand nombre de sens de mot en se basant sur le contexte (qui, dans notre cas, est la définition du dictionnaire), nous employons une modification de l'algorithme de Lesk. L'algorithme originel compare les mots pour savoir si ils appartiennent à un même lexème ou non; notre modification consiste en une comparaison floue employant un grand dictionnaire de synonyme et un système de morphologie dérivationnelle simple. L'application de la désambiguïsation aux définitions de dictionnaire (par contraste avec des textes habituels) permet quelques simplifications de l'algorithme (par exemple, nous ne nous soucions pas de la taille de la fenêtre de contexte).

Mots-clefs: *désambiguïsation du sens des mot, usage des ressources linguistiques, algorithme de Lesk*, *comparaison floue.*

We apply word sense disambiguation to the definitions in a Spanish explanatory dictionary. To calculate the scores of word senses basing on the context (which in our case is the dictionary definition), we use a modification of Lesk's algorithm. The algorithm relies on a comparison between two words. In the original Lesk's algorithm, the comparison is trivial: two words are either the same lexeme or not; our modification consists in fuzzy (weighted) comparison using a large synonym dictionary and a simple derivational morphology system. Application of disambiguation to dictionary definitions (in contrast to usual texts) allows for some simplifications of the algorithm, e.g., we do not have to care of context window size.

Keywords: *word sense disambiguation, usage of linguistic resources, Lesk's algorithm, fuzzy comparison*.

1 Introduction

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

We apply this improved algorithm to a large Spanish explanatory dictionary (30000 entries). This dictionary is not as "good" dictionary with restricted defining vocabulary as, for example, *Longman Dictionary of Contemporary 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 The main algorithm

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). For tagging, we use a set of syntactic heuristics developed specifically for this 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*).

Scoring. We represent each word sense as the set of words (except 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 Experimental results

We applied our algorithm to a Spanish dictionary of about 30000 entries, with the average number of words (except stopwords) per definitions being about 5. 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.

Let us look at two examples of the algorithm results. In the definitions of word senses we give in parenthesis the words that were used in scoring. The words between < > are the synonyms that were found, the words between [ ] are the words used in the definition of the word that is a part of the main definition. Between { } there are words that were scored because they were detected by the morphological model (substring comparison). Also we give the English translation of each Spanish phrase. The first example is for the word:

glándula = órgano que segrega sustancias indispensables para el organismo.
(*glandula = an organ that segregates indispensable substances for an organism*).

In this case the word *órgano* (*organ*) can have 4 meanings.

1. ([*sustancia*]) *Instrumento de viento, compuesto de un teclado y de un sistema de tubos por donde se emite el sonido*.
   (([*substance*]) *Wind instrument that is made of the clavier and a system of tubes which produces a sound.*)

2. ([*organismo*], [*organismo*], <*organismo*>,<*entidad*>) *Parte de un animal o planta que tiene una entidad en sí misma y realiza una función propia, como, por ejemplo, el estómago*.
   (([*organism*], [*organism*], <*organism*>,<*entity*>) *Part of an animal or a plant that is a unit itself and has its own function, like, for example, a stomach.*)

3. ([*organismo*],) *Cada una de las secciones del Estado, o de una empresa, con función propia*.
   (([*organism*]) *Each one of the sections of a state or an enterprise that has its own function*.)

4. ([*organismo*]) *Parte de una máquina que realiza una función específica*.
   (([*organism*]) *Part of a machine that has a specific function*.)

None of the words used in the word senses intersects directly with the main definition. Still the second sense got the greatest score because there are two words that have synonyms: *organism* and *entity* (namely, these words have each other as synonyms), besides, two words (*plant* and *animal*) have in their definitions the word *organism*. We use the latter fact basically to choose the correct sense in case of the equal scores, since we assign to it very small weight. In case of the first sense the word *tube* has in its definition the word *substance.* In the third and the fourth senses the word *function* has in its definition the word *organism.*

The other example is for a word:

operación = Negociación con valores bancarios.
(*operation = negotiation with banking values.*)

In this definition the word *valor* (*value*) can have one of the 11 meanings.

1. () *Precio, cualidad de las cosas por la que se paga cierta cantidad.*
   (() *Price, quality of things for which something is paid*.)

2. (<*valor*>) *Significado o importancia de algo dicho, escrito, etc*.
   ((<*value*>) *Significance, importance of something said, written, etc.*)

3. () *Cualidad del que no teme el peligro*.
   (() *Quality of do not be afraid of a danger.*)

4. () *Equivalencia, especialmente en monedas con respecto a las tomadas como patrón.*
   (() *Equivalence, especially of coins, with the respect to a standard one.*)

5. (<*valor*>,) *Grado de utilidad, importancia o buenas cualidades de algo.*
   ((<*value*>) *Degree of utility, importance or good qualities of something*.)

6. (<*valor*>,) *Atrevimiento, desvergüenza.*
   (<*courage*> *Insolence, impudence.*)

7. () *Firmeza, integridad*.
   (() *Firmness, integrity*.)

8. () *Eficacia.*
   (() *Efficiency.*)

9. () *Duración de una nota musical*.
   (() *Duration of a musical note.*)

10. ({*negoci*}, <*valor*>, <*valor*>) *Acciones, bonos o cualesquiera documentos negociables, acreditativos de una propiedad*.
    (({*negoti*}, <*value*>, <*value*>) *Shares, bonus or any negotiable documents, accrediting a property.*)

11. () *Persona que posee cualidades positivas para algo determinado.*
    (() *A person who has certain positive qualities for something determined*).

It can be seen that the sense which got the maximum score is the sense 10 because it has a morphological derivative (*negotiable* vs *negotiation*) and two words which have the word *value* as their synonym, namely *share* and *document*. The other senses have less scores.

4 Conclusions

We have suggested an improvement of Lesk's algorithm and applied it to a large Spanish explanatory dictionary. Our improvements consist in fuzzy comparison of words using a synonym dictionary and a simple derivational morphology procedure. Also, application of disambiguation to dictionary definitions (as compared with usual texts) allows for considerable simplification of the algorithm. Our algorithm gives better results than the original Lesk's algorithm and the baseline algorithm.

Acknowledgments

The work done under partial support of CONACyT, CGEPI-IPN, and SNI, Mexico.

References

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

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.

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

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.

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.