Tool for Computer-Aided Spanish Word Sense Disambiguation*

Yoel Ledo Mezquita[1, 2], Grigori Sidorov[1], and Alexander Gelbukh[1]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico D. F., Zacatenco, CP 07738, Mexico
{ledo, sidorov, gelbukh}@cic.ipn.mx, www.gelbukh.com

[2] Telematics Department, CUJAE, Cuba
ledo@tesla.ispjae.edu.cu

Abstract. We present a system for for computer-aided WSD mark-up of texts in Spanish. The system is is based on Anaya dictionary, uses a Spanish morphological analyzer and a WSD method based on Lesk algorithm (along with the other standard strategies). This tool reduces time and effort for preparation WSD-marked corpora in Spanish. We also discuss the requirement for such type of systems, which our particular system satisfies only partially.

1 Introduction

Words in a typical explanatory dictionary have different senses; this is known as polysemy. However, in a text each word occurrence corresponds to only one of these dictionary senses. The problem of determining this word sense used in a given text is referred to as word sense disambiguation (WSD). There are different methods for WSD that can be classified into two main groups: statistical methods [1, 4, 6, 10] and methods based on knowledge sources [3, 7, 8, 5].

The methods of either type require preliminary data preparation both for automatic learning and for automatic verification of results that permits to evaluate the quality of the method. Hence the necessity for a tool that would allow for manual or computer-aided sense marking in texts. We do not call it "semi-automatic" since the important decisions are taken by the human and not by the computer; an automatic or semiautomatic tool of this kind is currently impossible since modern WSD methods still have low precision.

In the rest of the paper, we first discuss the requirements for an "ideal" tool of this kind, and then describe the system we developed for Spanish, which satisfies the most part of these requirements.

* Work done under partial support of Mexican Government (CONACyT, SNI), IPN, Mexico (CGEPI, COFAA, PIFI), and RITOS-2.

A. Gelbukh (Ed.): CICLing 2003, LNCS 2588, pp. 277–280, 2003.

2 Requirements for a WSD Markup Tool

It is desirable that a system for computer-aided WSD markup of texts in any language be able to:

- Pass automatically to the next word in the text that can have different senses and present to the user a list of possible senses of each word (the words having only one sense can be marked automatically),

  In particular, the program should skip auxiliary words because their senses normally are irrelevant for WSD. However, the user should be able to manually choose the words that the program normally skips.

- Give the user a possibility to choose one or several senses that the word has in the given context with the minimum number of actions (clicks and movements),

- Suggest automatically the most probable sense(s) and then wait for a user confirmation.

  If the task is that multiple senses are allowed, then the confirmation is just a click on the OK button. However, if exactly one sense is to be chosen, then the user is to choose one sense from this small list; the senses should be ordered according to their probabilities so that in the majority of the cases the user could click at the OK button, which is equivalent to select the first one.

The system that skips auxiliary words and calculates the probabilities of word senses should use various procedures of linguistic analysis, namely:

- Processing of the given language's morphology:
  - Automatic morphological analysis,
  - Generation of lemmas,
  - Resolution of parts of speech (POS) ambiguity and ordering of lemmas according to the probabilities of their parts of speech in the text. If syntactic analysis (full or partial) is used for these purposes, then lemmas should be ordered according to the results of syntactic analysis.

- Implementation of different WSD strategies or their combinations (the user should have the possibility to choose the desired combination of these methods):
  - Statistical and/or knowledge-based methods. In addition, the option should be included to order the lemmas according to the POS probability, according to the WSD strategy, or some combination,
  - "Always first sense" strategy: the sense that is listed first in the dictionary is taken; this is rather good strategy because lexicographers tend to order senses intuitively according to their "importance" which in many cases coincides with their frequency in texts, see some considerations in [7],
  - "One sense per document" strategy [10]: the system supposes that the sense once used in the document will be repeated in the same document. For the first occurrence of the word, the sense is to be chosen by user and; for this, other strategies should be applied for suggesting the most probable sense, but all other occurrences of this word in the document are supposed to have the same

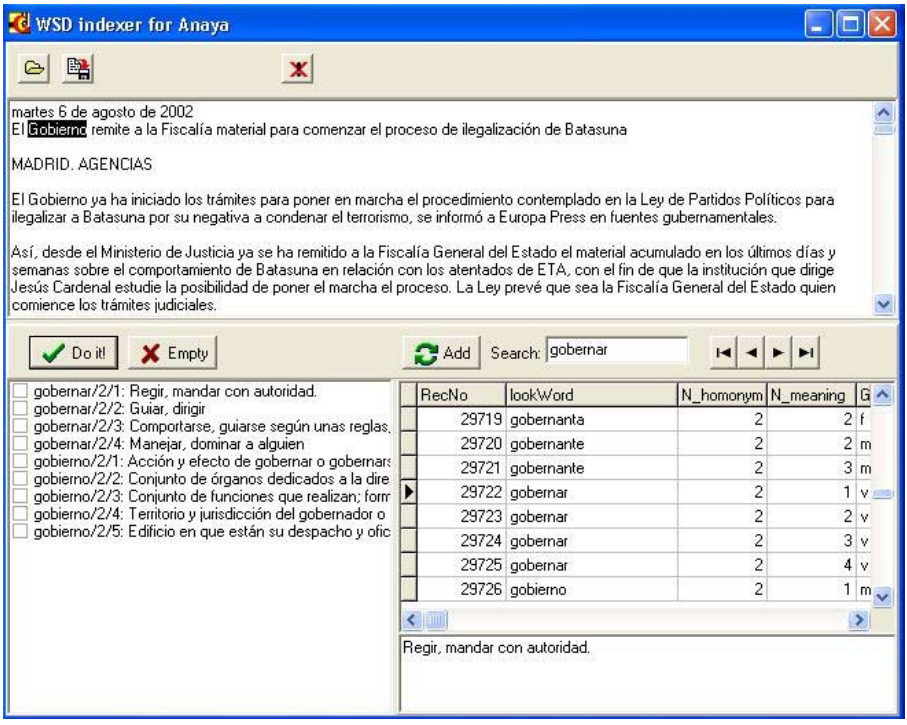

Fig. 1. Screenshot of the system.

sense. In addition, in this case the system should have an option of cleaning up all the data about the used word senses, in case if there are several documents in the file being processed.

We do not mention here some additional features: for example, what should be done in case that the dictionary is changed—say, some senses are merged or a new sense appears? The system should have a mode to reprocess the texts without unnecessary repetition of a manual work.

2 Tool We Developed

We developed such a tool for Spanish (Fig. 1). Of the requirements discussed above, the only one that our system does not implement is the resolution of POS ambiguity. Also, of the WSD strategies we have implemented only a version of the Lesk algorithm [7]; finally, we did not implement any graphical interface for combination of different WSD methods (this is changed directly in the program code if needed).

We used Anaya dictionary as the source for words and senses. This dictionary has more than 30,000 headwords. We preferred it over Spanish WordNet [9] because the latter has definitions in English while our WSD method needs definitions in Spanish.

It is possible to use any other explanatory dictionary in the corresponding format (we used a Paradox database).

For morphological processing, we applied a Spanish morphological analyzer / generator developed in our laboratory [1].

According to our experiments, the best results are achieved by combining the strategy of "one sense per document" and one of the WSD methods. The number of necessary clicks is more than 25% less in comparison with marking without system prompt. Note that the incorrect prompt is not penalized with any additional clicks because we use a mode in which only one sense is allowed.

3 Conclusions

We discussed the desired features of a system for computer-aided WSD marking of texts. We have presented a system for Spanish based on Anaya dictionary, which uses Spanish morphological analyzer and a WSD method based on the Lesk algorithm (along with some other standard strategies). The developed computer-aided tool allows for spending less time and effort for WSD text preparation in comparison with purely manual work.

References

1. Gelbukh, A. and G. Sidorov (2002). Morphological Analysis of Inflective Languages Through Generation. *J. Procesamiento de Lenguaje Natural*, No 29, September 2002, Spain. pp. 105–112.
2. Karov, Ya. and Edelman, Sh. (1998) Similarity-based word-sense disambiguation. *Computational linguistics*, Vol. 24, pp. 41–59.
3. 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.
4. Manning, C. D. and Shutze, H. (1999) *Foundations of statistical natural language processing*. Cambridge, MA, The MIT press, 680 p.
5. McRoy, S. (1992) Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, Vol. 18(1), pp. 1–30.
6. Pedersen, T. (2002) A baseline methodology for word sense disambiguation. In A. Gelbukh (ed.) "*Computational linguistics and intelligent text processing", LNCS 2276*, Springer, 2002, pp 126–135.
7. Sidorov G. and A. Gelbukh (2001). Word sense disambiguation in a Spanish explanatory dictionary. Proc. of *TALN-2001* (*Tratamiento automático de lengauje natural*), Tours, France, July 2–5, 2001, pp 398–402.
8. Wilks, Y. and Stevenson, M. (1999) Combining weak knowledge sources for sense disambiguation. Proceedings of *IJCAI-99*, 884–889.
9. *WordNet: an electronic lexical database*. (1998), C. Fellbaum (ed.), MIT, 423 p.
10. 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.