PAPEL: A lexical ontology for Portuguese

[Extended Abstract]

Hugo Gonçalo Oliveira
Linguateca, node of Coimbra,
DEI - FCTUC, CISUC,
Portugal
hroliv@dei.uc.pt

Diana Santos
Linguateca, node of Oslo,
SINTEF ICT, Norway
Diana.Santos@sintef.no

Paulo Gomes
Linguateca, node of Coimbra,
DEI - FCTUC, CISUC,
Portugal
pgomes@dei.uc.pt

Nuno Seco
Linguateca, node of Coimbra,
DEI - FCTUC, CISUC,
Portugal
nseco@dei.uc.pt

PAPEL (Palavras Associadas Porto Editora Linguateca) is a lexical resource for natural language processing (NLP) of Portuguese which is being built by Linguateca, based on processing a major commercial Portuguese dictionary, the Dicionário da Língua Portuguesa (DLP) [1] developed and owned by the largest Portuguese dictionary publisher, Porto Editora. There are known similar lexical resources for English and other languages. WordNet [10] is probably the most important reference and it is widely used in NLP research. WordNet was created from scratch, while we are creating the tools needed to build PAPEL (semi-)automatically. We could say that this work is mainly inspired by the MindNet [11] project. MindNet is a knowledge representation resource that used a broad-coverage parser to build semantic networks from machine readable human dictionaries (MRDs). The process of using MRDs to do NLP started back in the 1970's. Early known publications were from Nicoletta Calzolari [4]. A lot of work around MRDs took place during the 1980's ([9], [3], [5], [8], [2] and many more), but it was not until MindNet that we got an independent lexical ontology with knowledge automatically extracted from MRDs.

As far as we know, PAPEL is the first lexical ontology built by semi-automatic means for Portuguese, whose developments started October 2007.

To start our research some preliminary studies were made and some information about the dictionary structure was collected. After an empirical analysis of the most frequent ngrams, some initial grammars including specific string patterns to extract semantic relations were created and tested with a small portion of the dictionary (about 5000 entries). The grammars were developed for the PEN parser [1] which is a parser that implements the Earley's algorithm [7]. The first grammar extracted Cause-of/Caused-by relations between words. Currently we are working on grammars to extract the Part-of and Hypernymy relations. In the PAPEL project we have also developed a system for automatically identifying differences between the output of different versions of the same grammar, for regression testing.

In the future we intend to have a lexical resource whose base structures are concepts and not words. In order to accomplish our purpose we are hoping that, after having a network consisting of relations between words, some different senses of the same word might emerge. We are also planning is using some disambiguation methods (similar to [6]) to identify clustering related senses.

This presentation will start with a quick overview of existing similar resources, focusing on the the relations they provide. We will then show several patterns consisting of Portuguese words that can be used to identify some of those relations in the definitions of the DLP. Some examples of extracted relations and relations we believe we can extract will also be shown, as well as the tools developed. We will end with some discussion about the directions our work can take in the next months.

1. REFERENCES

[1] *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto, 2005.

[2] Alshawi, H. Processing dictionary definitions with phrasal pattern hierarchies. *Comput. Linguist. 13*, 3-4 (1987), 195–202.

[3] Amsler, R. A. A taxonomy for english nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 1981), Association for Computational Linguistics, pp. 133–138.

[4] Calzolari, N. An empirical approach to circularity in dictionary definitions. In *Cahiers de Lexicologie* (1977), pp. 118–128.

[5] Chodorow, M. S., Byrd, R. J., and Heidorn,

[1] http://linguateca.dei.uc.pt/index.php?sep=recursos

G. E. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 1985), Association for Computational Linguistics, pp. 299–304.

[6] Dolan, W. B. Word sense ambiguation: clustering related senses. In *Proceedings of the 15th conference on Computational linguistics* (Morristown, NJ, USA, 1994), Association for Computational Linguistics, pp. 712–716.

[7] Jurafsky, D., and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[8] Markowitz, J., Ahlswede, T., and Evens, M. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics* (Morristown, NJ, USA, 1986), Association for Computational Linguistics, pp. 112–119.

[9] Michiels, A., Mullenders, J., and Noël, J. Exploiting a large data base by longman. In *Proceedings of the 8th conference on Computational linguistics* (Morristown, NJ, USA, 1980), Association for Computational Linguistics, pp. 374–382.

[10] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. Introduction to wordnet: An on-line lexical database*. *Int J Lexicography 3*, 4 (January 1990), 235–244.

[11] Richardson, S. D., Dolan, W. B., and Vanderwende, L. Mindnet: Acquiring and structuring semantic information from text. In *COLING-ACL* (1998), pp. 1098–1102.