PAPEL: A Dictionary-Based Lexical Ontology for Portuguese

Hugo Gonçalo Oliveira[1], Diana Santos[2], Paulo Gomes[1], and Nuno Seco[1]

[1] Linguateca, Coimbra node, DEI - FCTUC, CISUC, Portugal
[2] Linguateca, Oslo node, SINTEF ICT, Norway
hroliv@dei.uc.pt, Diana.Santos@sintef.no, pgomes@dei.uc.pt, nseco@dei.uc.pt

Abstract. This paper describes a project aimed at creating a lexical ontology extracted (semi) automatically from a large Portuguese general dictionary. Although using machine readable dictionaries to extract semantic information is not new, we believe this is the first attempt for the Portuguese language. The paper describes a (to be) freely available resource, dubbed PAPEL, explaining the process used and the tools developed, and illustrating it with one specific relation: Causation.

1 Introduction

PAPEL (Palavras Associadas Porto Editora Linguateca) is a lexical resource for natural language processing (NLP) of Portuguese, based on the (semi) automatic extraction of relations between the words appearing in the definitions of a general language dictionary of Portuguese - the Dicionário da Língua Portuguesa (DLP) [1] developed and owned by the largest Portuguese dictionary publisher, Porto Editora. Similar lexical resources for English are Princeton WordNet [2] widely used by NLP researchers, and MindNet [3]. When it comes to Portuguese, despite projects with similar aims (WordNet.PT [4] and WordNet.BR [5]), there is no publicly available lexical ontology for our language (i.e., that one can download and use in its entirety).

Also, and differently from the two aforementioned projects, which are done from scratch resorting to extensive manual linguistic labour, we follow the approach of creating PAPEL from a machine readable dictionary (MRD).

This paper starts with a description of the most important works that, since the 1970's, have used MRDs as a source of information to solve the lexical bottleneck in NLP, pointing out the similarities of PAPEL compared to these earlier attempts. It then describes, in Section 3, the methodology employed in the creation of PAPEL, and the tools developed in the project. Section 4 explores in some detail the example of Causation, while Section 5 ends with a description of some further work.

2 Related Work

This section presents an overview on the most important works that used MRDs as a source of information for NLP, especially for the extraction of relations

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 31–40, 2008.

between words or concepts and the creation of organized structures containing those relations. Similar resources are then described.

2.1 Relation Extraction from MRDs

The process of using MRDs in natural language processing (NLP) started back in the 1970's, with the work of Nicoletta Calzolari [6] [7]. Definitions were explored in order to organize the dictionary into a lexical database where morphological and semantic information about the defined words could be obtained directly. Similar work took place for English when the electronic versions of the Longman Dictionary of Contemporary English (LDOCE) and the Merriam-Webster Pocket Dictionary (MPD) were used as a source of information to build such a structure. The analysis of the structure of those MRDs showed that they made use of a very limited defining vocabulary [8] and that the text of the definitions often consisted of a *genus* an a *differentia* [9]. The *genus* identifies the superordinate concept of the defined word. The *differentia* presents the properties responsible for the distinction between this "instance" of the superordinate concept and other instances of the same concept. Amsler [10] suggested that the identification of the *genus* could lead to the construction of a taxonomy. Bearing in mind the definition structure, Chodorow [11] took advantage of its restricted vocabulary and developed semi-automatic recursive procedures to extract and organize semantic information into hierarchies. These heuristics didn't need to parse the whole definitions, due to their predictability. However, the human user played an important role when it came to disambiguation. Other approaches [12] [13] took advantage of the simple vocabulary of the definitions and used string patterns to extract semantic information from them.

Further approaches [14] [15] used patterns based on the structural level (i.e., syntactic phrases) of the analysed text, instead of string patterns. After some discussion about the advantages and the drawbacks of using structural patterns or string patterns to extract semantic information contained in the definitions, Montemagni and Vanderwende [16] concluded that although string patterns are very accurate for identifying the *genus*, they cannot capture the variations in the *differentia* as well as structural patterns, and they proposed the use of a broad-coverage grammar to parse the dictionary definitions in order to obtain rich semantic information. In spite of seeming an overkill to use a broad-coverage parser for definition text, the authors make the point that in some cases (relative clauses, parenthetical expressions, and coordination) its use is warranted. Although dictionaries have been explored for several purposes, such as parsing or word sense disambiguation, to our knowledge they have not been converted into an independent resource of its own before MindNet [3], which therefore can be said to be a sort of independent lexical ontology in a way that previous work was not.

2.2 Related Resources

Princeton WordNet [2] is probably the most important reference when it comes to lexical ontologies in English. It is freely available and it is widely used in NLP

research. In the WordNet's lexicon, the words are clearly divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the *synset*, which is a set of synonym words that can be used to represent one concept. The *synsets* are organized as a network of semantic relations, such as Hyponymy and Meronymy (between nouns) and Troponymy and Entailment (between verbs).

WordNet.BR [5] is a Brazilian version of the "wordnet concept", started in 2002. Their database is structured around Synonymy and Antonymy manually extracted from a reference corpus where several dictionaries are included, and plans for adding more relations in the future have been reported in [5]. WordNet.PT [4] is another attempt of creating a Portuguese lexical resource from scratch, which started in 1999. The authors of WordNet.PT explicitly claim that the available resources for Portuguese NLP are not suitable for the automatic construction of such a resource. They use a set of 35 relations and are explicitly interested in cross-categorical relations such as those linking adjectives to nouns.

MindNet [17] is a knowledge representation resource that used a broad-coverage parser to build a semantic network, not only from MRDs but also from encyclopedias, and free text. MindNet contains a long set of relations, including Hypernymy, Causation, Meronymy, Manner, Location and many more. One interesting functionality offered by MindNet is the identification of "relation paths" between words[1]. For example, if one looks for paths between *car* and *wheel* a long list of relations will be returned. The returned paths include not only simple relations like *car is a modifier of wheel* but also more complex ones like *car is a hypernym of vehicle and wheel is a part of vehicle.*

Another kind of lexical resource is FrameNet [18], which constitutes a network of relations between semantic frames, extracted from corpora and from a systematic analysis of semantic patterns in corpora. Each frame corresponds to a concept and describes an object, a state or an event by means of syntactic and semantic relations of the lexical item that represents that concept. A frame can be conceived as the description of a situation with properties, participants and/or conceptual roles. A typical example of a semantic frame is *transportation*, within the domain *motion*, which provides the elements *mover(s)*, *means of transportation* and *paths* and can be described in one sentence as: *mover(s) move along path by means.*

3 Building PAPEL

In this section, we describe the set of relations included in PAPEL, the parser used to analyse the definitions, some quantitative studies about the content of the definitions, the incremental nature of the work and the regression testing tools developed in this project.

3.1 Relations

The overview of the resources referred in Section 2.2, together with an exploration of the most common n-grams in the dictionary, led us to choose the first

[1] http://atom.research.microsoft.com/mnex/

set of relations that we want to have in PAPEL. Note that the decision of working on a relation means also the detection of its inverse.

Let us start by explaining that the most basic semantic relationship between words is, of course, identity of meaning (**Synonymy**, and in fact *synsets* in wordnets are simply a set of words having the same meaning), but we started by assuming that other semantic relations would be more interesting for general NLP applications and that their discovery would facilitate the identification of the set of final concepts. This is related to the often made remark that word sense disambiguation is an ill-defined task and is very dependent on the purpose [19]. Different lexicographers, or system developers, divide senses differently [20]. So we consider the task of ambiguating a dictionary [21] a task more germane to our interests than word sense disambiguation.

Table 1 shows some of the relations we are planning to include in PAPEL, their representation and some examples. These relations include the *is-a* relation (**HIPONIMO_DE**), the causation relation (**CAUSADOR_DE**) and the *part-of* relation (**PARTE_DE**).

Table 1. Relations we are planning to include in PAPEL

Relation	**Inverse**	**Example**
HIPERONIMO_DE(X,Y)	HIPONIMO_DE(Y,X)	**HIPERONIMO_DE**(animal, cão)
CAUSADOR_DE(X,Y)	RESULTADO_DE(Y,X)	**CAUSADOR_DE**(vírus, doença)
PARTE_DE(X,Y)	INCLUI(Y,X)	**PARTE_DE**(roda, carro)
MEIO_PARA(X,Y)	FINALIDADE_DE(Y,X)	**MEIO_PARA**(chave, abrir)
LOCAL_DE(X,Y)	OCORRE_EM(Y,X)	**LOCAL_DE**(restaurante, comer)

We are also planning to deal with other kinds of relations that should be easy to extract and that we thought would considerably increase the usefulness of the resource are words related to places (`lisboeta` related to `Lisboa`) and words describing affect (positive or negative connotation).

3.2 Parsing the Definitions

In order to parse the dictionary definitions, we used PEN, a chart parser freely available under a BSD license[2] which is a Java implementation of the well known Earley Algorithm [22]. PEN parses the text according to a grammar file it gets as input and it can yield several analysis for the same text. So far, we have used specific different grammars to identify different relations between the defined entities corresponding to words in the dictionary.

The relation extraction method starts with an empirical analysis of the patterns present in the definitions and which might suggest relations between the entry and other entities. Having a relation in mind, a selection of patterns (e.g. `tipo de X`) that can imply the relation is made.

An SQL table containing information about the n-grams in the definitions of the dictionary was created, providing us with the frequency of each n-gram

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

in the whole dictionary, its position inside the definition, and the grammatical category of the defined word. Guided by the frequency of the candidate patterns in the definitions, we look at a selection of entries where the patterns are actually used to make sure their selection makes sense and to possibly find more refined criteria as well.

After finding a set of patterns indicating a relation, we can start the construction of a specific grammar for the extraction of that relation in the dictionary.

To deal sensibly with multiple analyses of a same definition according to the same relation, we implemented the following heuristic in every grammar: the selected derivation is the one with less unknown tokens.

3.3 The Results

After having devising and debugging the grammars with the help of a set of hand-selected definitions (about 5000), we apply them to the whole dictionary, comprising 237,246 definitions.

We then analyse the results for the whole dictionary in order to classify the relations obtained into "correct" and "incorrect". This classification is made by a human user and can be very time consuming. That is why we have created a program to automate part of the process. We can feed the program with a set of correct and a set of incorrect relations from previous runs. The human user then only has to classify the relations which are not in any of the previous sets, making time spent to obtain the first division pay off in the following runs of new versions of the grammar(s) for the same relation.

In fact, the number of **new** kinds of problems drastically diminishes as more relations are classified, because since the dictionary definitions use simple and not very diverse vocabulary (though not as restricted as LDOCE), most of the problems detected are systematic (see Section 4.2 for examples of obtained errors). The number of "correct" and "incorrect" candidate relations extracted give us an idea of when to stop developing further the grammars.

3.4 Regression Testing

After analysing the relations considered correct and the incorrect ones, it is easier to find out the origin of the problems. This analysis helps us deciding what changes should be made in the grammar. The new version of the grammar is then tested, before processing the whole dictionary.

The results obtained with different versions of a grammar for the extraction of the same relation can be compared with a system we have developed especially for regression testing. This system identifies differences between two sets of results and can be used to obtain information about, and quantify:

- the relations in one set and not in the other;
- the relations that remained the same in both sets;
- the entries that have at least one relation in one set but any in the other;
- the changes to the relations obtained for each entry;

4 Detailed Example: Causation

We proceed by describing in some detail the process and results obtained for **CAUSADOR_DE** relation, namely defined by us as a relation between an agent (the causer) and a result (the caused). We have considered the inverse relation, **RESULTADO_DE**, to be the same as effect/result, taking thus so far a naive approach to this philosophical debate (see e.g. [23] or [24]).

As described above, we developed several grammars to parse the dictionary definitions that included these relations, and went on testing them incrementally. When it comes to this relation, we currently have a 96% success rate (precision) in a total of 5,657 **CAUSADOR_DE** relations extracted and 91% in a total of 1,693 **RESULTADO_DE** relations. These numbers were calculated after manual analysis of the results. We are starting to look into corpus-based methods to evaluate recall.

4.1 The Patterns

The grammars designed for the extraction of this relation are primarily based on the verbs `causar`, `originar`, `provocar`, `produzir`, `motivar`, `gerar`, `suscitar` and `resultar` and on the expressions `devido a` and `efeito de`.

The following patterns are used for the extraction of the **CAUSADOR_DE** relation.

```
1c - causad{o|a|os|as} FREQ* PREP CAUSADOR
     originad{o|a|os|as} FREQ* PREP CAUSADOR
     provocad{o|a|os|as} FREQ* PREP CAUSADOR
     produzid{o|a|os|as} FREQ* PREP CAUSADOR
     gerad{o|a|os|as} FREQ* PREP CAUSADOR
     motivad{o|a|os|as} FREQ* PREP CAUSADOR
     suscitad{o|a|os|as} FREQ* PREP CAUSADOR
2c - devido {a|ao|à|aos|às} CAUSADOR
3c - efeito PREP CAUSADOR
```

`CAUSADOR` is a sub-pattern that denotes a **CAUSADOR_DE** relation between words it catches (which will be the cause) and the entry (which will be the result): **CAUSADOR_DE**(cause, entry). The cause can be preceded by specific words like determiners, pronouns, quantifiers, other modifiers or constructions like `acção de/do/dos/da/das`.

`PREP` denotes a preposition and `FREQ` a (optional) quantifier, such as `normal mente` or `frequentemente`.

The following patterns are used for the extraction of the **RESULTADO_DE** relation:

```
1r - que {causa|origina|provoca|produz|motiva|gera|suscita} RESULTADO
2r - {causar|originar|provocar|produzir|motivar|gerar|suscitar} RESULTADO
3r - resultar PREP_EM RESULTADO
```

The sub-pattern `RESULTADO` is similar to `CAUSADOR`, but catches the results in the definition instead of catching the causes: **RESULTADO_DE**(result, entry).

Table 2. Examples of relations extracted by the grammars for the Causation grammars. 'ID' identifies the pattern matched by the definition.

ID	Entry	Grammar	Definition	Extracted relation
1c	quase-delito	s. m.	dano **causado** por negligência, sem intenção malévola	**CAUSADOR_DE**(negligência, quase-delito)
1c	concussão	s. f.	choque violento **originado** por uma explosão	**CAUSADOR_DE**(explosão, concussão)
1c	toxicose	s. f.	doença **provocada** pela presença de produtos tóxicos no organismo	**CAUSADOR_DE**(produtos ,toxicose)
1c	ecfonema	s. m.	elevação súbita da voz, **motivada** por surpresa ou comoção violenta	**CAUSADOR_DE**([surpresa, comoção], ecfonema)
1c	tisne	s. m.	cor **produzida** pelo fogo ou pelo fumo sobre a pele	**CAUSADOR_DE**([fogo, fumo], tisne)
2c	engasgo	s. m.	incapacidade de respirar devido a obstrução da garganta	**CAUSADOR_DE**(obstrução, engasgo)
3c	maximização[3]	s. f.	**efeito de** maximizar	**CAUSADOR_DE**(maximizar, maximização)
1r	diplodoco	s. f.	bactéria que **causa** as meningites cerebrospinais	**RESULTADO_DE**(meningites, diplodoco)
1r	osteoporose	s. f	porosidade excessiva dos ossos, que **origina** a sua fragilidade	**RESULTADO_DE**(fragilidade, osteoporose)
1r	tentação	s. f.	coisa ou pessoa que **provoca** desejo	**RESULTADO_DE**(desejo, tentação)
2r	penalizar[3]	v. tr.	**causar** pena, dor, aflição a	**RESULTADO_DE**([pena, dor, aflição], penalizar)
2r	inimizar[3]	v. tr.	**provocar** inimizade entre	**RESULTADO_DE**(inimizade, inimizar)
3r	displasia	s. f.	desenvolvimento anormal de um órgão ou de um tecido, de que podem **resultar** deformidades graves	**RESULTADO_DE**(deformidades, displasia)

In pattern 3c `PREP_EM` denotes the preposition `em` contracted or not with a determiner.

Note that we also deal with enumeration of causes or effects/results separated by commas or conjunctions using a recursive rule that overcomes the "conjoined heads" problem, which is one of the limitations of using string patterns pointed by [16].

4.2 Results

Table 2 shows some examples of the relations extracted by the Causation grammars.

Manual inspection of the obtained relations yielded 4% erros in **CAUSADOR_DE** and 8% in **RESULTADO_DE** relations. Examples of the most common errors are:

[3] Note that these patterns discover relations between nouns, as well as between nouns and verbs, which may probably be better modelled by other relation names such as ACCAO_QUE_CAUSA and RESULTADO_DA_ACCAO.

1. definitions that mention the relation between two words of the definition, and not relative to the entry word: `estetoscópio, s. m. - instrumento para auscultar a respiração, as batidas do coração e outros sons produzidos pelo corpo`, **CAUSADOR_DE**(corpo, estetoscópio);
2. definitions where the pattern is preceded by a negative word, making the entity a "non-cause": `respeitar, v. tr. - não causar dano`, **RESULTADO_DE**(dano, respeitar);
3. definitions using brackets: `inspirar, v. tr. - provocar (ideias, pensamentos, projectos)`, **RESULTADO_DE**((, inspirar);
4. definitions using commas: `heterocarpo, adj. - que produz, espontaneamente ou por intervenção do homem, flores ou frutos diferentes`, **RESULTADO_DE**(, , heterocarpo).

Items 1 and 3 are pointed out by [16] as limitations of using string patterns instead of structural patterns to extract relations from text.

5 Conclusions and Further Work

This project intends to create a computationally tractable ontology from mining a particular (general language) dictionary, and not provide THE ontology for Portuguese. In further (separate) projects we might investigate overlap with other sources for ontology (other dictionaries, reference works, corpora etc.) but this is outside the scope of PAPEL. So, corpus-based validation of PAPEL is simply a way of detecting further patterns in the dictionary to add rules for the particular relations, and not any general corpus-based ontology creation.

We are doing improvements to PEN in order to be able to decouple morphological and lexical information from the grammar. In this respect, we intend to try out a broad-coverage parser such as PALAVRAS [25].

We are also devising a system to help humans revising the residual examples that are not amenable to automatic parsing, so that they will be easily included in the final resource and possibly also feed the dictionary proper.

After the extraction of the relations, we will have a network of words linked by relations. We are considering the hypothesis of performing a process similar to the one described in [21] to identify groups of related definitions inside the same entry (word) and use them for the ultimate construction/detection of synonyms and *synsets*.

Acknowledgments

We would like to thank the group of R&D of Porto Editora for making their dictionary available for this research. The project PAPEL is supported by the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC.

References

1. Dicionário PRO da Língua Portuguesa. Porto Editora, Porto (2005)
2. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An on-line lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
3. Richardson, S.D., Dolan, W.B., Vanderwende, L.: Mindnet: Acquiring and structuring semantic information from text. In: COLING-ACL, pp. 1098–1102 (1998)
4. Marrafa, P., Amaro, R., Chaves, R.P., Lourosa, S., Martins, C., Mendes, S.: Wordnet.pt new directions. In: Sojka, P., Choi, K.-S., Fellbaum, C., Vossen, P. (eds.) Proceedings of GWC 2006: 3rd International Wordnet Conference, Jeju Island, Korea, pp. 319–320 (2006)
5. Dias-da-Silva, B.C.: Wordnet.br: An exercise of human language technology research. In: Sojka, P., Key-Sun Choi, C.F., Vossen, P. (eds.) Proceedings of the Third International WordNet Conference — GWC 2006, South Jeju Island, Korea, January 22–26 (2006)
6. Calzolari, N.: An empirical approach to circularity in dictionary definitions. In: Cahiers de Lexicologie, pp. 118–128 (1977)
7. Calzolari, N.: Detecting patterns in a lexical data base. In: Proceedings of the 22nd annual meeting on Association for Computational Linguistics, pp. 170–173. Association for Computational Linguistics, Morristown (1984)
8. Michiels, A., Mullenders, J., Noël, J.: Exploiting a large data base by Longman. In: Proceedings of the 8th conference on Computational linguistics, pp. 374–382. Association for Computational Linguistics, Morristown (1980)
9. Amsler, R.A.: The structure of the Merriam-Webster Pocket dictionary. PhD thesis, The University of Texas at Austin (1980)
10. Amsler, R.A.: A taxonomy for english nouns and verbs. In: Proceedings of the 19th annual meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics, Morristown (1981)
11. Chodorow, M.S., Byrd, R.J., Heidorn, G.E.: Extracting semantic hierarchies from a large on-line dictionary. In: Proceedings of the 23rd annual meeting on Association for Computational Linguistics, pp. 299–304. Association for Computational Linguistics, Morristown (1985)
12. Markowitz, J., Ahlswede, T., Evens, M.: Semantically significant patterns in dictionary definitions. In: Proceedings of the 24th annual meeting on Association for Computational Linguistics, pp. 112–119. Association for Computational Linguistics, Morristown (1986)
13. Alshawi, H.: Analysing the dictionary definitions. Computational lexicography for natural language processing, 153–169 (1989)
14. Vanderwende, L.: Algorithm for automatic interpretation of noun sequences. In: Proceedings of the 15th conference on Computational linguistics, pp. 782–788. Association for Computational Linguistics, Morristown (1994)
15. Vanderwende, L.: Ambiguity in the acquisition of lexical information. In: Proceedings of the AAAI 1995 Spring Symposium Series, pp. 174–179 (1995); symposium on representation and acquisition of lexical knowledge
16. Montemagni, S., Vanderwende, L.: Structural patterns vs. string patterns for extracting semantic information from dictionaries. In: Proceedings of the 14th conference on Computational linguistics, pp. 546–552. Association for Computational Linguistics, Morristown (1992)

17. Vanderwende, L., Kacmarcik, G., Suzuki, H., Menezes, A.: Mindnet: An automatically-created lexical resource. In: HLT/EMNLP. The Association for Computational Linguistics (2005)
18. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th international conference on Computational linguistics, pp. 86–90. Association for Computational Linguistics (1998)
19. Wilks, Y.: Is word sense disambiguation just one more nlp task? Computers and the Humanities 34, 235–243 (2000)
20. Kilgarriff, A.: Word senses are not bona fide objects: implications for cognitive science, formal semantics, nlp. In: Proceedings of the 5th International Conference on the Cognitive Science of Natural Language Processing, Dublin, pp. 193–200 (1996)
21. Dolan, W.B.: Word sense ambiguation: clustering related senses. In: Proceedings of the 15th conference on Computational linguistics, pp. 712–716. Association for Computational Linguistics, Morristown (1994)
22. Earley, J.: An efficient context-free parsing algorithm. Communications of the ACM 6(8), 451–455 (1970)
23. Vendler, Z.: Causal relations. The Journal of Philosophy 64, 704–713 (1967)
24. Vendler, Z.: Effects, results and consequences. Linguistics in Philosophy 64, 147–171 (1967)
25. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Protuguese in a Constraint Grammar Framework. PhD thesis, Arhus University, Arhus (2000)