Extracting semantic relations from Portuguese *corpora* using lexical-syntactic patterns

Raquel Amaro
Group for the Computation of Lexical and Grammatical Knowledge
Center of Linguistics of the University of Lisbon
Avenida Prof. Gama Pinto, 2, 1649-003 Lisbon – Portugal
E-mail: ramaro@clul.ul.pt

Abstract

The growing investment on automatic extraction procedures, together with the need for extensive resources, makes semi-automatic construction a new viable and efficient strategy for developing of language resources, combining accuracy, size, coverage and applicability. These assumptions motivated the work depicted in this paper, aiming at the establishment and use of lexical-syntactic patterns for extracting semantic relations for Portuguese from *corpora*, part of a larger ongoing project for the semi-automatic extension of WordNet.PT. 26 lexical-syntactic patterns were established, covering hypernymy/hyponymy and holonymy/meronymy relations between nominal items, and over 34 000 contexts were manually analyzed to evaluate the productivity of each pattern. The set of patterns and respective examples are given, as well as data concerning the extraction of relations - right hits, wrong hits and related hits-, as well as the total of occurrences of each pattern in CPRC. Although language-dependent, and thus clearly of obvious interest for the development of lexical resources for Portuguese, the results depicted in this paper are also expected to be helpful as a basis for the establishment of patterns for related languages such as Spanish, Catalan, French or Italian.

Keywords: lexical-syntactic patterns, semantic relations, *corpora*

1. Introduction

The advances in the area of Natural Language Processing (NLP) make apparent the importance of understanding and processing informational content of natural languages, critical for the development of an ever growing number of applications dealing with word sense disambiguation, information retrieval, machine translation, human-machine communication, and so on. In this context, lexical semantic resources, specifically ontological or concept-based relational ones, play a crucial role and their development constitutes a major concern for researchers in the fields of Computer Science and Computational Linguistics (Jing *et al.* 2000; Wandmacher *et al.* 2007). Also in this context, wordnets (lexical-conceptual relational databases; Miller 1990, Fellbaum 1998, Vossen 2002) gain importance and usability, being their development widely stimulated.

Given that most NLP applications require lexicons of 20 to 60000 word-forms (Dorr & Jones 1996), the need for extensive resources has led to many fully automatic approaches to extract ontological knowledge from raw or structured data (Maedche 2002, Cimiano & Staab 2005), resulting however in large but unreliable resources, as expected in unsupervised procedures. On the contrary, fully manual developed lexica are very reliable, but time-consuming, expensive and often small. The two strategies have been followed in the development of wordnet-like resources for Portuguese: manual construction, as WordNet.PT[1], and automatic construction, whether by the translation of or the strict alignment with already existent wordnets, as MultiWordNet[2], whether by the extraction of semantic relations from corpora, thesauri or dictionaries, as Onto.PT[3]. The translation and alignment strategies can be suitable alternatives for some applications but do not solve issues concerning different lexicalizations of concepts (Hirst 2004), for instance, and are limited to the input data and relations and subject to options and mistakes of a third party, non-speaker of the target language. Fully automatic extraction of lexica and semantic relations from data, on the other hand, can sometimes cope with language-dependent phenomena, but deals with inaccurate degrees of semantic granularity, ambiguity and parsing issues and inference inconsistencies (Oliveira & Gomes 2010). Finally, manually built lexica result in accurate, balanced and dense resources, but too often of non-sufficient size.

The growing investment on automatic extraction procedures, together with the already mentioned need for extensive resources, makes semi-automatic construction a new viable and efficient strategy for developing such resources, which combine accuracy, size, coverage and applicability (Piasecki *et al.* 2009). These assumptions motivated the work depicted in this paper, aiming at the establishment and use of lexical-syntactic patterns for extracting semantic relations for Portuguese from *corpora*, part of a larger ongoing project for the semi-automatic extension of WordNet.PT.

Although language-dependent, and thus clearly of obvious interest for the development of lexical resources for Portuguese, the results depicted in this paper are expected to be helpful as a basis for the establishment of

[1] http://www.clul.ul.pt/clg/wordnetpt/index.html

[2] http://mwnpt.di.fc.ul.pt/index.html

[3] http://ontopt.dei.uc.pt/

patterns for related languages such as Spanish, Catalan, French or Italian.

2. General approach

The determination of lexical-syntactical patterns conveying semantic relations aims at extracting candidates semantically and conceptually related from *corpora*, as opposed to structured data such as dictionaries or encyclopedias. The following sections describe the data used and the methodology followed to obtain the relevant patterns.

2.1 Data description

Aiming at defining and testing lexical-syntactical patterns in language data for the purpose of extracting semantic relations, more specifically, lexical items related by hypernymy/hyponymy and holonymy/meronymy relations, the data used was the Reference Corpus of Contemporary Portuguese[4] (CRPC), developed by the Center of Linguistics of the University of Lisbon.

CRPC is a large electronic *corpus* of more than 311 million words, mostly from European Portuguese. It covers several types of written texts (309.8 million words) and spoken texts (1.6 million words), dated from the second half of the 19th century up until 2006, although the majority of texts are dated from after 1970. The written texts include literary, newspaper, technical, scientific, didactic, leaflets, decisions of the supreme court of justice, parliament sessions texts, etc. The spoken sub *corpus* includes formal and informal speech of a variety of spoken interaction types: monologues, dialogues, conversations, phone conversations, lectures and homilies. CRPC can be characterized as a 'reference' *corpus* since its composing texts are sampled before being included in the corpus.

The lexical-syntactical patterns described in this paper were tested through the CRPC online queries tool.

2.1 Methodology

2.1.1 Pattern drafting

The determination of patterns expressing semantic relations – to extract candidates semantically and conceptually related from raw data – considered

a) distributional information, based on the assumption that items similar in meaning tend to occur in the same contexts (Fillmore 1968; Grimshaw 1990; Jackendoff 1990; Hearst 1991; Levin 1993), and collocation analysis;

b) lexical-syntactic patterns (Hearst 1992), handmade patterns defined *a priori* according to the target language structures; also considering syntactic and morph and syntactic information, such as recurrent modification structures (Buitelaar et al. 2004) or predicate-argument relations (Wagner 2000),

to assure the wider coverage possible. This resulted in two types of patterns, according to the way they were designed: handmade (HM) patterns and distributional information-based (DI) patterns.

HM patterns were first shaped based on the relations available in WordNet.PT and on the linguistic tests defined for each relation, an approach that gathers several advantages (Piasecki et al. 2009: 105), but also on introspective language knowledge. The candidate patterns were then tested in controlled data (CRPC v2.3 mainly, but also in non-controlled data through online queries for low frequency items).

The results of these patterns were manually revised and analyzed in order to determine which patterns worked and which did not. Not always the results were the expected: the pattern did occur in the *corpus* or on in non-controlled data; the pattern did not extract directly the semantic relation aimed at.

The patterns that returned results were then registered in a database to be further used; the patterns that did not worked were abandoned.

DI patterns emerged from the analysis of collocations and distributional information in CRPC of particular lexical items already encoded in WordNet.PT, i.e. items linked by a particular semantic relation: a given lexical item was search for in CRPC, collocations were obtained from these results and then analyzed to find semantically related words; or a given pair of semantic related words was looked for in CRPC. This analysis allowed for determining more patterns relating disambiguated word-forms and possible candidates.

This preliminary stage did not include automatic acquisition of patterns, considered for a later stage of the project. The future approach for the automatic acquisition of patterns requires a first testable set of patterns, needed to evaluate the automatic acquisition results. In this way, the results of the current work function also as the basis for future automatic acquisition of patterns.

The work depicted in this paper focuses on nouns, since these constitute the majority of items in the database. However, data on several cross-part of speech (POS) relations concerning other POS, including adverbs and prepositions, was collected for future analysis, specifically in what concerns function (agent, patient, and so on) or other less salient relations in wordnet model, such as cause or manner relations.

2.1.2 Pattern adjustment

The HM and DI candidate patterns obtained were subject to a screening process since their determination was not always straightforward.

In what concerns HM patterns, not always the selected lexical items produced results: i) they did not occur in the data; ii) the related items did not co-occur in the same sentence or in the considered 5 to 10 word pre and post target interval; and so different formulations had to be tested or the pattern was excluded.

Also, sometimes the results were not straightforward: the pattern did not occur; the pattern did not always only express the particular semantic relation tested; and in the last case, not always the semantic relation tested was the most prominent relation expressed by the pattern.

More than 50 patterns were first drafted and tested, resulting

[4] http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc

in the definition of 26 patterns with potential use for extracting candidates from *corpora*. Ambiguous, rare or wrong patterns were dismissed.

2.1.3 Pattern search and selection

The next step involved the large-scale use of both HM and DI patterns on CRPC data to extract lexical candidates related by semantic relations. Each of the patterns was converted into a query regular expression to extract the contexts of occurrence in CRPC, using, whenever possible and productive, POS annotation tags to boost the search accuracy. All results, in a total of 34 747 contexts, were manually analyzed to accurately evaluate the productivity of the pattern, i.e., the number of occurrences that expressed a given relation obtained *vs*. the non-relevant occurrences and, thus, the potential of usability of the pattern in the semi-automatic extension of WordNet.PT.
Results are presented in the next section.

3. Results

The lexical-syntactic patterns established cover hypernymy/hyponymy and holonymy/meronymy relations between nominal items. Table 1 presents this set of patterns, with examples from CRPC.

Relation: hyperonymy (N_1 is hyperonym of N_2)			
Nr.	**Pattern**	**Example**	**Type**
P1	∅ N_1, *especialmente/nomeadamente* N_2 N_1 especially/namely N_2	***lacticínios***, *especialmente* ***manteiga*** **dairy products**, especially **butter**	DI
P2	DET N_1, *especialmente/nomeadamente* DET N_2 DET N_1, especially/namely DET N_2	*os* ***resíduos*** *, nomeadamente a* ***biomassa*** the **residues**, namely the **biomass**	DI
P3	PREP$_1$ DET N_1, *especialmente/nomeadamente* PREP$_1$ DET N_2 PREP$_1$ DET N_1, especially/namely PREP$_1$ DET N_2	*pelos* ***insectos***, *especialmente pelas* ***moscas*** by the **insects**, especially by the **flies**	DI
P4	N_1, *como* N_2, N_1, as N_2	***cancros***, *como* ***leucemia*** **cancers**, as **leucemia**	DI
P5	N_1s,*tais como* N_2, N_1s, such as N_2	*diversos* ***desportos***, *tais como* ***natação***, several **sports**, such as **swimming**,	HM
Relation: hyponymy (N_1 is hyponym of N_2)			
P6	DET.IND N_1*é* DET.IND N_2 DET.IND $N_1$is DET.IND N_2	*um* ***isótopo*** *é um* ***átomo*** an **isotope** is an **atom**	HM
P7	DET.DEF N_1*é* DET.IND N_2 DET.DEF $N_1$is DET.IND N_2	*o* ***óleo*** *é um* ***resíduo*** the **oil** is a **residue**	HM
P8	PONTUAÇÃO *os/as* N_1s *são* N_2s PUNCTUATION the N_1s are N_2s	*os* ***computadores*** *são* ***máquinas*** the **computers** are **machines**	HM
P9	PONTUAÇÃO ∅ N_1s *são* N_2s PUNCTUATION ∅ N_1s are N_2s	***creches*** *são* ***estabelecimentos*** **daycare centers** are **facilities**	HM
P10	N_1 *é um tipo de* N_2 N_1 is a kind/type of N_2	***pedofilia*** *é um tipo de* ***crime*** **pedophilia** is a type of **crime**	HM
P11	N_1 *e/ou outro/as* N_2s N_1 and/or other N_2s	***escolas*** *ou outras* ***instituições*** **schools** or other **institutions**	DI
P12	*apontar|apresentar|definir|descrever|identificar|indicar* DET.DEF N_1 *como* DET.DEF N_2 point out|present|define|describe|identify|indicate DET.DEF N_1 as DET.DEF N_2	*apontando o* ***catolicismo*** *como a* ***religião*** pointing out the **catholicism** as the **religion**	DI
P13	*classificar|conceber|considerar|entender|qualificar|tomar* DET.DEF N_1 *como* DET.DEF N_2 classify|conceive|consider|figure|qualify|take DET.DEF N_1 as DET.DEF N_2	*se tomasse a* ***azálea*** *como a* ***flor*** if (he) took the **azalea** as the **flower**	DI
P14	*aludir|confirmar|consagrar|lembrar|reconhecer* DET.DEF N_1 *como* DET.DEF N_2 mention|confirm|acclaim|remember|recognize DET.DEF N_1 as DET.DEF N_2	*confirmar o* ***azeite*** *como a* ***gordura*** confirm the **olive oil** as the **fat**	DI
P15	*estabelecer|impor|instituir|proclamar* DET.DEF N_1 *como* DET.DEF N_2 establish|impose|institute|proclame DET.DEF N_1 as DET.DEF N_2	*estabelece a* ***segunda-feira*** *como o* ***dia*** establish the **monday** as the **day**	DI

Relation: holonymy (N_1 is holomym ofN_2)			
P16	*dividir* DET N_1 *em* N_2s divide DET N_1 in DET N_2s	*dividindo o* ***livro*** *em* ***capítulos*** dividing the **book** in **chapters**	HM
P17	N_1 *ser formado/a*(*s*) *por* (DET) N_2s N_1 be formed/composed by (DET) N_2s	*uma* ***família*** *é formada por* ***indivíduos*** a **family** is formed by **individuals**	HM
P18	N_1 *ser feito/a*(*s*) *de* (DET) N_2s N_1 be made of (DET) N_2s	*o* ***preservativo*** *é feito de* ***látex*** the **condom** is made of **latex**	HM
P19	N_1 *ser constituído/a(s) por* (DET) N_2s N_1 be composed by (DET) N_2s	*uma* ***reserva*** *ser constituída por* ***terrenos*** a **reservation** be constituted by **lands**	HM
P20	DET N_1 *é um conjunto*|*grupo* de N_2 DET N_1 is a set|group of N_2	*uma* ***espécie*** *é um conjunto de* ***seres*** a **species** is a set of **beings**;	HM
P21	N_1 *ser feito com* N_2 N_1 be made with N_2s	*a* ***marinada*** *é feita com* ***alho*** *, louro* the **marinade** is made with **garlic**, bay leaf	HM
Relation: meronymy (N_1 is meronym ofN_2)			
P22	N_1 *ser a parte de* (DET) N_2s N_1 be the part of (DET) N_2s	*a* ***metafísica*** *é a parte da* ***filosofia*** the **metaphysics** is the part of **philosophy**	HM
P23	*reparação de* (DET) N_1 *de* (DET) N_2 repair of (DET) N_1 of (DET) N_2	*reparação da* ***pista*** *do* ***aeroporto*** repair of the **runway** of the **airport**	DI
P24	*reparar*|*consertar* DET.DEF N_1 *de* (DET.DEF) N_2 repair|fix DET.DEF N_1 of (DET.DEF) N_2	*reparar a* ***fachada*** *da* ***casa*** repair the **façade** of the **house**	DI
P25	N_1*é* DET *zona de*+DET N_2 N_1 is DET area of+DET N_2	***mácula****, que é a zona da* ***retina*** ***macula****,that is an area of the* ***retina***	HM
P26	N_1 *é* DET *espaço de*+DET N_2 N_1 is DET space of+DET N_2	*a* ***cozinha*** *é o espaço central da vida da* ***casa*** the **kitchen** is the central space of+the life of+the **house**	HM

Table 1: PT lexical-syntactic patterns for nominal hypernymy/hyponymy and holonymy/meronymy relations

	Right hits	*Related hits*	*Wrong hits*	*Total occurrences*
P1	36.2%	9.3%	54.5%	789
P2	53.8%	28.7%	17.5%	383
P3	50.5%	11.9%	37.6%	1021
P4	33.6%	5.7%	58.7%	506
P5	65.1%	4.8%	30.2%	63
P6	26.1%	0.2%	73.7%	1063
P7	24.8%	n.a.	75.2%	10914
P8	18.9%	0.9%	80.2%	1740
P9	11.7%	2.2%	86.1%	231
P10	33.3%	3.7%	63.0%	54
P11	58.4%	0.6%	41.0%	14874
P12	17.4%	9.9%	72.7%	161
P13	24.5%	9.5%	66.0%	200
P14	29.4%	20.6%	50.0%	34
P15	66.7%	n.a	44.3%	18
P16	21.7%	n.a.	78.3%	143
P17	31.2%	8.7%	60.1%	173
P18	12.2%	0.6%	87.2%	500
P19	20.6%	11.1%	68.3%	848
P20	42.6%	10.3%	47.1%	68
P21	12.7%	n.a.	87.3%	379
P22	32.0%	n.a.	68.0%	25
P23	7.8%	n.a.	92.2%	377
P24	16.6%	n.a.	84.0%	95
P25	20.7%	12.1%	67.2%	58
P26	30.0%	3.3%	66.7%	30

Table 2: Quantitative results of patterns search

Data concerning relations extraction - right hits, wrong hits and related hits, as well as the total of occurrences of each pattern in CPRC, are presented in Table 2.Related hits concern cases of i) indirect relations, for instance, the lexical items retrieved are and indirectly related (ex.: *escravo* – *indivíduo* (slave – individual), instead of *escravo* – *pessoa* (– *indivíduo*) (slave – person (– individual))); ii) proper names (ex.: *jornalistas* – *Ramirez* (journalists – Ramirez)); iii) items associated by relations different from the targeted one (ex.: *cabine* – *casa* (cabin – house)).

Patterns whose related hits were higher than the right hits (concerning the targeted relation) were dismissed for a given relation or, according to the analyzed results, used to extract other relations.

The percentage of wrong hits is variable, being quite high in several patterns. However, in most cases, right hits do not amount to the majority of cases, even in "paradigmatic" patterns such as "is a", for hyponymy, (pattern P6) or "is part of", for meronymy (pattern P22) that do not perform as efficiently as could be expected. These results further motivate the need for human intervention to assure accuracy.

For these reasons, and at this stage of the project, the extracted relations are always submitted to manual revision. The final process considers human decision, to include or exclude automatically extracted relations in the WordNet.PT database, central.

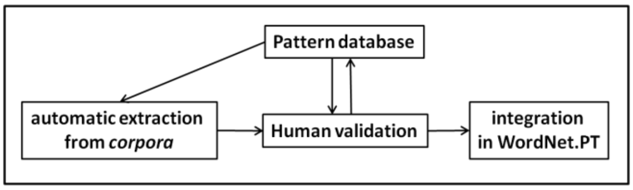
Pattern database
automatic extraction from *corpora*
Human validation
integration in WordNet.PT

Figure 1: Semi-automatic extension of WordNet.PT process

However, and to minimize the human effort in this process, several strategies for automatic selection of results are being considered, based on the results analyzed and considering frequent expressions (locutions and other structures) that fit the pattern and that can be dismissed before human validation.

4. Final remarks

The establishment of lexical-syntactic patterns for Portuguese has as main goal the extraction of semantic relations from *corpora*, as part of a larger ongoing project aiming at the semi-automatic extension of WordNet.PT.

As illustrated above, a necessary element of this project concerns, thus, a reliable pattern database to provide the grounds to anchor the whole process. For this reason, the resulting database of patterns will be publicly released, pending project conclusion.

The lack of this type of data for Portuguese, on the one hand, and the importance of this kind of language-dependent information for the automatic and/or manual development of lexical resources on the other, led us to consider that the publication and availability of such a database will be of great value for research and for language resources development for Portuguese.

However, although language-dependent, the results depicted in this paper are expected to be also helpful as a basis for the establishment of patterns for related languages such as Spanish, Catalan, French or Italian.

5. Acknowledgements

The work presented here has been supported by the post-doctoral fellowship SFRH/BPD/75904/2011 granted by Fundação para a Ciência e a Tecnologia. The author also wishes to thank the remarks from the anonymous reviewers that contributed to the final version of this paper.

6. References

C Buitelaar, P., Olejnik, D., Hutanu, M., Schutz, A., Declerck, T. and Sintek, M. (2004). Towards ontology engineering based on linguistic analysis. In *Proceedings of the Lexical Resources and Evaluation Conference*. Lisbon, Portugal, pp. 7--10.

Cimiano, P. and Staab, A. (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In C. Biemann & G. Paas (Eds.), *Proceedings of the ICML Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*. Bonn, Germany.

Dorr, B.J. and Jones, D. (1996). Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In E. Viegas (Ed.), *Breadth and Depth of Semantic Lexicons*. Association of Computational Linguistics, pp. 42--50.

Fellbaum, C. (1998). A Semantic Network of English: the Mother of all WordNets. In P. Vossen (Ed.), *EuroWordNet: a Multiligual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, pp. 137--148.

Fillmore, C.J. (1968). The Case for Case. In E. Bach & R.T. Harms (Eds.), *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, pp. 1--88.

Grimshaw, J. (1990). *Argument Structure*. Cambridge: The MIT Press.

Hearst, M.A (1991). Noun Homograph Disambiguation Using Local Context in Large Text Corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*. Oxford, England, pp. 1--22.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France, pp. 539--545.

Hirst, G. (2004). Ontology and the lexicon. In S. Staab & R. Studer (Eds), *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, pp. 209--230.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Jing, H., Dahan Netzer, Y., Elhadad, M. and McKeown, K.R. (2000). Integrating a large-scale, reusable lexicon with a natural language generator. In *Proceedings of the 1st International Conference on Natural Language Generation*. Mitzpe Ramon, Israel, pp. 209--216

Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation*. Chicago: IL.

Maedche, A. (2002). *Ontology learning for the semantic web*. Kluwer Academic Publishers.

Miller, G. A. (1990). *WordNet: an online Lexical Database*. Special Issue of International Journal of Lexicography, 3(4).

Oliveira, H. and Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In T. Agnote (Ed.) *Stairs 2010: Proceedings of 5th European Starting AI Researcher Symposium*. IOS Press, pp. 199--211.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Wrocław, Poland: Oficyna Wydawnicza Politechniki Wrocławskiej.

Vossen, P. (Ed.) (2002). *EuroWordNet General Document*. EuroWordNet Project LE2-4003 & LE4-8328 report. University of Amsterdam.

Wagner, A. (2000). *Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis*. In *Proceedings of the ECAI: Workshop on Ontology Learning*. Berlin, Germany.