Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese

Hugo Gonçalo Oliveira[1] **Paulo Gomes**[2]

Abstract. This ongoing research presents an alternative to the manual creation of lexical resources and proposes an approach towards the automatic construction of a lexical ontology for Portuguese. Textual sources are exploited in order to obtain a lexical network based on terms and, after clustering and mapping, a wordnet-like lexical ontology is created. At the end of the paper, current results are shown.

1 INTRODUCTION

In the last decade, besides the increasing amount of Semantic Web [2] applications, we have seen a growing number of systems that perform tasks where understanding the information conveyed by natural language plays an important role. Natural language processing (NLP) tasks, from machine translation or automatic text generation to intelligent search, are becoming more and more common, which demands better access to semantic knowledge.

Knowledge about words and their meanings is structured in lexical ontologies, such as Princeton WordNet [15], which are used in the achievement of the aforementioned tasks. Since this kind of resource is most of the times handcrafted, its creation and maintenance involves time-consuming human effort. So, its automatic construction from text arises as an alternative, providing less intensive labour, easier maintenance and allowing for higher coverage, as a trade-off for lower, but still acceptable, correction.

This paper presents Onto.PT, an ongoing research project where textual resources, more precisely dictionaries, thesaurus and corpora, are being exploited in order to extract lexico-semantic knowledge that will be used in the construction of a public domain lexical ontology for Portuguese. While the first stage of this work deals mainly with information extraction from text, subsequent stages are concerned with the disambiguation of the acquired information and the construction of a structure similar to WordNet. Considering that information is extracted from different sources, one particular point is that we aim to accomplish word sense disambiguation (WSD) [28] based not on the context where information is found but on knowledge already extracted. Therefore, clustering over extracted synonymy instances is first used to identify groups of synonymous words that will be used as a conceptual base. The rest of the information, consisting of term-based triples, is then mapped to the conceptual base as each term is assigned to a group of synonyms.

After introducing some background concepts and relevant work, we state the goals of this research. Then, we introduce the stages involved in the approach we are following. Before concluding, current results of this project, as well as their evaluation, when available, are shown.

2 BACKGROUND KNOWLEDGE

Besides recognising words, their structure and their interactions, applications that deal with information in natural language need to understand its meaning, which is usually achieved with the help of knowledge bases assembling lexical and semantic information, such as lexical ontologies. Despite some terminological issues, lexical ontologies can be seen both as a lexicon and as an ontology [22], and are significantly different from classic ontologies — they are not constructed for a specific domain and are intended to provide knowledge structured on lexical items (words) of a language by relating them according to their meanings. In this context, Princeton WordNet [15] is the most representative lexico-semantic resource for English and also the most common model for representing a lexical ontology. WordNet is structured on synsets, which are groups of synonymous words describing concepts, and connections, denoting semantic relations (e.g. hyponymy, part-of), between those groups.

The success of the WordNet model led to its adoption by many lexical resources in several different languages, such as the wordnets involved in the EuroWordNet [38] project, or WordNet.PT [26], for Portuguese. However, the creation of a wordnet, as well as the creation of most ontologies, is typically manual, thus involving much human effort [4]. To overcome this problem, some authors [9] propose the translation of a target wordnet to wordnets in other languages. This seems to be a suitable alternative for several applications but another problem arises because different languages represent different socio-cultural realities, they do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently [22]. Another popular alternative for ontology creation is to extract lexico-semantic knowledge and learn lexical ontologies from text, which can either be unstructured, as in textual corpora, or semi-structured, as in dictionaries or encyclopedias.

Research on the acquisition of lexico-semantic knowledge from corpora is not new and varied methods, roughly divided into linguistics-based (see [20, 32]), statistics or graph-based (see [36, 25, 14]) or hybrid (see [6, 7, 1, 18, 17]), have been proposed to achieve different steps of this task, such as the extraction of relations like hyponymy [20, 6, 7], meronymy [1, 17], causation [18], or the establishment of sets of similar or synonymous words [32, 25, 36].

Dictionary processing, which became popular during the 1970s [5], is also a good option for the extraction of this kind of knowledge. MindNet [31] is both an extraction methodology and a lexical ontology different from a wordnet, since it was created automatically from a dictionary and its structure is based on such resources. Nevertheless, it still connects sense records with semantic relations (e.g. hyponymy, cause, manner). Most of the research on the automatic creation of lexical resources from electronic dictionaries was made

[1] PhD student, CISUC, University of Coimbra, Portugal, hroliv@dei.uc.pt

[2] CISUC, University of Coimbra, Portugal, pgomes@dei.uc.pt

during the 1980s and 1990s, where the advantages and drawbacks of using the later resources were studied and discussed [23]. Still, there are reports of recent works on the automatic extraction of knowledge from dictionaries (see [29, 19, 27]). For instance, PAPEL [19] is a lexical resource consisting of a set of triples denoting semantic relations between words found in a Portuguese dictionary.

Besides corpora and dictionary processing, in the later years, semi-structured collaborative resources such as Wikipedia or Wiktionary, have proved to be important sources of lexico-semantic information and have thus been receiving more and more attention by the research community (see for instance [33, 21, 40, 27]).

On the one hand, there are clear advantages of using dictionaries — they are already structured on words and meanings, they cover the whole language, and they generally use simple and almost predictable vocabulary. On the other hand, dictionaries are static resources with limited knowledge. Therefore, some authors [20, 32] argue that textual corpora should be exploited to extract knowledge that can be found neither in dictionaries nor in lexical ontologies. Also, while language dictionaries are not always available for this kind of research, there is always much text available on the most different subjects, for instance in the Web. The biggest problem concerning lexico-semantic information extraction from corpora is that there are no boundaries on the vocabulary and linguistic constructions used, thus leading to more ambiguity and parsing issues.

Most of the aforementioned works on the extraction of semantic relations from text output related words, identified by their orthographical form. However, since natural language is ambiguous, this representation is not practical for most computational applications, because the same orthographical form might either have completely different meanings (e.g. *bank*, institution or slope) or closely related meanings (e.g. *bank*, institution or building). Furthermore, there are words with completely different orthographical forms denoting the same concept (e.g. *car* and *automobile*). This might lead to serious inconsistencies, for instance when dealing with inference, as in an example, in Portuguese, reported in [19]: *queda* SYNONYM_OF *ruína* $\wedge$ *queda* SYNONYM_OF *habilidade* $\rightarrow$ *ruína* SYNONYM_OF *habilidade*. Here, the two words in the inferred relation are almost opposites and not synonyms.

Therefore, another challenge on lexical ontology learning from text, often called ontologising, is concerned with moving from knowledge based on words to knowledge based on concepts. For English, there are works on the assignment of suitable WordNet synsets to the arguments of relational triples extracted from text, or to other term entities, such as Wikipedia entries [33]. Some of the methods for ontologising term-based triples compute the similarity between the context from where each triple was extracted with the terms in synsets, sibling synsets or direct hyponym synsets [35]. Others look for relations established with the argument terms and with the terms of each synset [30], or take advantage of generalisation through hypernymy links [30].

3 RESEARCH GOALS

The main goal of this research is the automatic construction of Onto.PT, a broad-coverage structure of Portuguese words according to their meanings, or, more precisely, a lexical ontology.

Regarding information sparsity, it seems natural trying to create such a resource with knowledge extracted from several sources, as proposed in [39] for creating a lexical ontology for German, but for Portuguese. Thus, we are using or planning to use the following sources of knowledge: (i) dictionaries, such as Dicionário da Língua Portuguesa [13], through PAPEL; Dicionário Aberto (DA) [34], an open domain electronic version of a Portuguese dictionary from 1913; and the Portuguese Wikitionary[3]; (ii) encyclopedias, such as the complete entries of the Portuguese Wikipedia[4] or just their abstracts; (iii) corpora, yet to decide; and (iv) thesaurus, such as TeP [12], an electronic thesaurus for Brasilian Portuguese; and OpenThesaurus.PT[5], a thesaurus for European Portuguese.

Considering each resource specificities, such as its organisation or the vocabulary used, the extraction procedures might be significantly different, but must have one common output: a set of term-based relational triples. Still, considering the limitations of representations based on the terms, we are adopting a wordnet-like structure which enables the establishment of unambiguous semantic relations between synsets. Moving from a lexical network to a lexical ontology requires the application of several WSD techniques. However, our intention is to achieve WSD based only on knowledge already extracted, because we believe this is the best way to harmoniously integrate knowledge coming from different heterogeneous sources. Another point that should be considered is the attribution of confidence weight(s) on each relation, based on its frequency and also on one or several similarity measures, calculated according to the words distribution in a corpus.

4 PROPOSED APPROACH

In this section, we describe all the stages involved in the creation of Onto.PT, also represented in Figure 1. Furthermore, we give an overview on possible ways to evaluate the results of this work, which, in the future, will be freely available.

4.1 Extraction of relational triples

The first stage on the creation of Onto.PT is the automatic extraction of lexico-semantic knowledge from textual sources. The extracted information is represented as relational triples, $t_1\ R\ t_2$, where t_1 and t_2 are terms and R is the name of a semantic relation held between possible meanings of t_1 and t_2. These triples establish a lexical network, $L = (N, E)$, with $|N|$ nodes and $|E|$ edges, $E \subset N^2$, where each node $i \in N$ is a term and each edge between nodes i and j, $E(i, R, j)$, means that a relation of the type R between nodes i and j was extracted.

Hence, each sentence is analysed by a parser according to semantic grammars created specifically for each relation to be extracted. Most of the rules in the semantic grammars are based on textual patterns frequently used to denote each semantic relation, such as the ones presented in Table 1 for well-known relations in Portuguese.

Table 1. Examples of patterns indicating semantic relations.

Relation	Example pattern
Hypernymy	`tipo|género|classe|forma de`
Meronymy	`parte|membro de`
Causation	`causado|provocado|originado por`
Purpose	`usado|utilizado|serve para`

Extraction from dictionaries follows very closely the extraction procedure described in [19]. Despite significant differences in dictionary and corpora text, the general extraction procedure works for

[3] http://pt.wiktionary.org

[4] http://pt.wikipedia.org

[5] http://openthesaurus.caixamagica.pt/

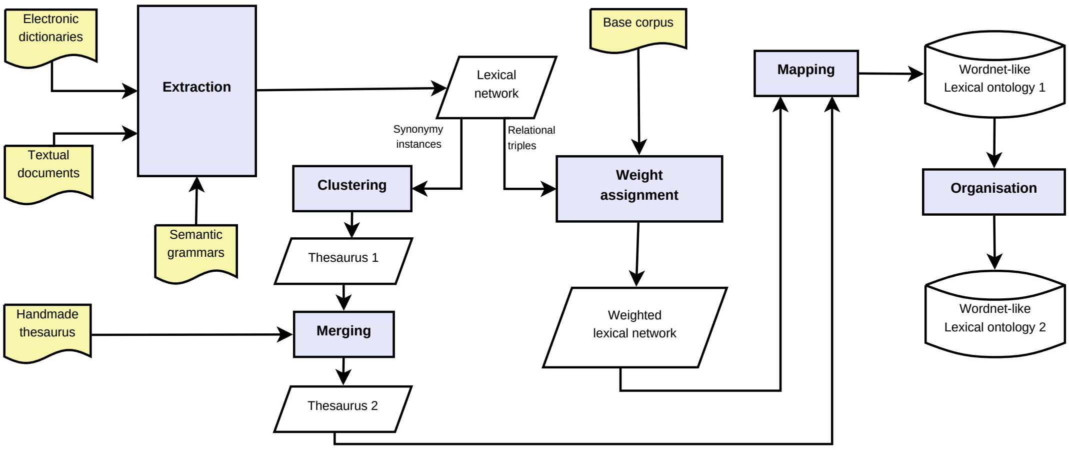

Figure 1. Information flow in the construction of Onto.PT

both, with slightly differences in the construction of the grammars. For instance, most of the relations extracted from dictionary definitions are established between a word in the definition and the word being defined. Moreover, dictionaries are important to obtain synonymy instances (t_1 SYNONYM_OF t_2), since many words are defined by (a list of) their synonyms. On the other hand, despite sharing most of the times the same neighbourhood, synonymy instances may not co-occur frequently in corpora text [14].

In order to find less intuitive patterns, a pattern discovery algorithm [20] can be applied over a corpus: (i) a relation R is chosen; (ii) several pairs of words known to establish R are looked for in a corpus; (iii) everytime both words of the same pair co-occur in a sentence, the text connecting them is collected; (iv) most frequent sentences collected are used as hints for new patterns denoting R.

4.2 Clustering for synsets

Since lexical resources based on the words orthographical form are inadequate to deal with ambiguity, we are adopting a wordnet-like structure, where concepts are described by synsets and ambiguous words are included in a synset for each of their meanings. Semantic relations can thereby be unambiguously established between two synsets, and concepts, even though described by groups of words, will bring together natural language and knowledge engineering in a suitable representation, for instance, for the Semantic Web. Moreover, this makes it possible to apply inference rules for the discovery of new knowledge.

From a linguistic point of view, word senses are complex and overlapping structures [24, 22]. So, despite word sense divisions in dictionaries and ontologies being most of the times artificial, this trade-off is needed in order to increase the usability of broad-coverage computational lexical resources.

As lexical synonymy networks extracted from dictionaries tend to have a clustered structure [16], clusters are identified in order to establish synsets. A possible way to achieve clustering and deal with ambiguity at the same time, is to use a hard-clustering algorithm, such as the Markov Clustering Algorithm (MCL) [37], and extend it to find unstable nodes, which are most of the times ambiguous words. This is the approach of [16], that runs clustering with noise several times, creates a matrix with the probabilities of each node belonging to each cluster, and finally, assigns each word to all the clusters its belonging probability is higher than a threshold.

Our procedure is very similar and is described as follows: (i) split the original network into sub-networks, such that there is no path between two elements in different sub-networks, and calculate the frequency-weighted adjacency matrix F of each sub-network; (ii) add stochastic noise to each entry of F, $F_{ij} = F_{ij} + F_{ij} * \delta$; (iii) run MCL over F for 30 times; (iv) use the (hard) clustering obtained by each one of the 30 runs to create a new matrix P with the probabilities of each pair of words in F belonging to the same cluster; (v) create the clusters based on P and on a given threshold $t = 0.2$. If $P_{ij} > t$, i and j belong to the same cluster; (vi) in order to clean the results, remove: (a) big clusters, B, if there is a group of clusters $C = C_1, C_2, ...C_n$ such that $B = C_1 \cup C_2 \cup ... \cup C_n$; (b) clusters completely included in other clusters.

4.3 Merging with other synset-based resources

In this stage, we take advantage of broad-coverage synset-based resources for Portuguese, such as thesaurus, in order to enrich our synset base. Still, we are more interested in manually created resources of that kind, since they can amplify the coverage, and improve the precision of our synsets at a significantly low cost.

The following procedure is applied for merging two thesaurus: (i) define one thesaurus as the basis B and the other as T; (ii) create a new empty thesaurus M and copy all the synsets in B to M; (iii) for each synset $T_i \in T$, find the synsets $B_i \in B$ with higher Jaccard coefficient[6] c, and add them to a set of synsets $J \subset B$. (iv) considering c and J, do one of the following: (a) if $c = 1$, it means that the synset is already in M, so nothing is done; (b) if $c = 0$, T_i is copied to M; (c) if $|J| = 1$, remove J_1 from M and add a new synset $N = J_1 \cup T_i$ to M. (d) if $|J| > 1$, a new set, $N = T_i \cup J'$ where $J' = \cup_{i=0}^{|J|} J_i$, $J_i \in J$, is added to M and all synsets in J are removed from M.

4.4 Assigning weights to triples

In this stage, one (or several) weights are assigned to triples based on the number of times they were extracted (frequency) and also on distributional metrics, calculated over a corpus. The later metrics,

[6] $Jaccard(A, B) = A \cap B / A \cup B$

typically used to retrieve similar documents, assume that similar or related words tend to co-occur or to occur in similar contexts. Nevertheless, several distributional metrics (e.g. latent semantic analysis (LSA) [11]) have also been adapted to measure the similarity of two words, based on their neighbourhoods [7, 39]. The weights can thus be used to indicate the confidence for each triple and thresholds can be applied to discard lower-weighted triples and improve precision. For instance, [8] reports high correlations between manual evaluation of hypernymy and part-of triples and their weights according to some distributional measures computed on a corpus.

4.5 Mapping term-based triples to synsets

After the previous stages, a thesaurus T and a term-based lexical network, L, are available. In order to set up a wordnet, this stage uses the latter to map term-based triples to synset-based triples, or, in other words, assign each term, a and b, in each triple, $(a\ R\ b) \in L$, to suitable synsets of T. This task, often called ontologising [30], can be seen as WSD, but we explicitly aim to achieve disambiguation by taking advantage of knowledge already extracted, and not of the context from where it was extracted. Having this in mind, we have developed two mapping methods.

In the first method, to assign a to a synset A, b is fixed and all the synsets containing a, $S_a \subset T$, are obtained. If a is not in T, it is assigned to a new synset $A = (a)$. Otherwise, for each synset $S_{ai} \in S_a$, n_{ai} is the number of terms $t \in S_{ai}$ such that $(t\ R\ b)$ holds. Then, the proportion $p_{ai} = \frac{n_{ai}}{|S_{ai}|}$ is calculated. All the synsets with the highest p_{ai} establish a set C. Finally, (i) if $|C| = 1$, a is assigned to the only synset in C; (ii) if $|C| > 1$, C' is the set of elements of C with the highest n_a and, if $|C'| = 1$, a is assigned the synset in C', unless $p_{ai} < \theta$ [7]; (iii) if it is not possible to assign a synset to a, it remains unassigned. Term b is assigned to a synset using this procedure, but fixing a. In a second phase, we take advantage of hypernymy links already established to help mapping semi-mapped triples, which are triples where one of the arguments is assigned to a synset and the other is not ($A\ R\ b$ or $a\ R\ B$).

The second mapping method starts by creating a term-term matrix, M, based on the adjacencies of the lexical network. Consequently, M is a square matrix with n lines, where n is the total number of nodes (terms) in the lexical network. If the term in index i and the term in index j are connected by some kind of relation, $M_{ij} = 1$, otherwise, $M_{ij} = 0$. In order to assign synsets to a and b, the first thing to do is, once again, to get all the synsets including the term a, $S_a \subset T$, and also all synsets including b, $S_b \subset T$. Then, the similarity between each synset $A \in S_a$ and each synset $B \in S_b$ is given by the average lexical network based similarity for each term in A with each term in B:

$$sim(A, B) = \frac{\sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \cos(A_i, B_j)}{|A||B|}$$

Here, the similarity of two vectors of M gives us the similarity of two words, based on their neighbourhoods in the lexical network, and is calculated by the cosine of their adjacency vectors, A_i and B_j respectively. To conclude the mapping, the pair of synsets with a higher similarity is chosen.

4.6 Knowledge organisation

In this stage, routines for knowledge organisation are applied in order both to make it possible to infer new implicit knowledge and also to remove redundant triples. This is achieved by applying some rules to the synset-based triple set, including:

- Transitivity: if R is transitive (e.g. SYNONYMY, HYPERNYMY, ...), $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$
- Inheritance: if R is not a HYPERNYMY or HYPONYMY relation, $(A \text{ HYPERNYM\_OF } B) \wedge (A\ R\ C) \rightarrow (B\ R\ C)$

Therefore, some behaving properties, such as transitivity, inheritance or inversion, of the extracted relations are predefined. For instance, to deal with inversion, all relations are only stored in the type defined as the direct one, but, if needed, the system can inverse them.

4.7 Evaluation

Evaluation takes place through all the previous stages. Manual evaluation is a reliable kind of evaluation, but it is also time-consuming and difficult to reproduce, so, when possible, we are willing to explore automatic evaluation procedures. Automatic evaluation is typically performed by comparing the results obtained with a gold standard, but the later is not always available, especially for a broad-coverage ontology, where freely available gold standards are scarce.

The validation of relational triples can also be perfomed using a collection of documents to find hints on them. For instance, triples can be translated to common natural language patterns, such as the ones in Table 1, and looked for in that form, as in [10] to assign probabilities to semantic triples, or in [19], to validate them.

Moreover, the quality of the final ontology will also be assessed when using it to perform NLP tasks, such as question answering or automatic generation of text.

5 CURRENT RESULTS

Since the authors of this research are also part of the PAPEL development team, PAPEL can be seen as a seed project. So, in Table 2, we start by presenting the numbers and examples of some of the relations included in PAPEL 2.0, and also the numbers of the relations obtained after applying exactly same extraction procedure, described in [19], to DA. We have taken advantage of the grammatical information provided by the dictionaries to organise each type of relation according to the grammatical category of its arguments.

Table 2. Relations extracted from dictionaries.

Relation	Arguments	PAPEL 2.0	DA	Examples
Synonymy	noun,noun	37,452	20,910	*auxílio, contributo*
	verb,verb	21,465	8,715	*tributar, colectar*
	adj,adj	19,073	7,353	*flexível, moldável*
	adv,adv	1,171	605	*após, seguidamente*
Hypernymy	noun,noun	62,591	59,887	*planta, salva*
Part-of	noun,noun	2,805	1,795	*cauda, cometa*
	noun,adj	3.721	4,902	*tampa, coberto*
Member-of	noun,noun	5.929	1,564	*ervilha, Leguminosas*
	adj,noun	883	59	*celular, célula*
Causation	noun,noun	1.013	264	*fricção, assadura*
	adj,noun	498	166	*reactivo, reacção*
	verb,noun	6,399	5,714	*limpar, purgação*
Purpose	noun,noun	2,886	1,760	*defesa, armadura*
	verb,noun	5,192	3,383	*fazer_rir, comédia*
	verb,adj	260	186	*corrigir, correccional*

The relations between nouns in a previous version of PAPEL were validated (also in [19]), by searching for natural language sentences

[7] θ is a threshold defined to avoid that a is assigned to a big synset where a, itself, is the only term related to b.

denoting the relations in a newspaper corpus. About 20% of the part-of and hypernymy triples were supported by the corpus. On the other hand, these numbers were respectively 10% and 4% for purpose and causation. The results are interesting since there is not as much general knowledge in a newspaper as in a dictionary and because we have used a small set of patterns when there is a huge amount of possibilities to denote these semantic relations in corpora text.

Table 3. Relations extracted from Wikipedia abstracts.

Relation	Quant.	Example	Sample	Corr.	Agr.
Synonymy	11,862	*estupro,violação*	286	86,1%	91,2%
Hypernymy	29,563	*estilo_de_música,folk*	322	59,1%	93,1%
Part-of	1,287	*jejuno,intestino*	268	52,6%	78,4%
Causation	520	*parasita,doença*	244	49,6%	79,5%
Purpose	743	*construção,terracota*	264	57,0%	82,2%

Moving on to other kinds of text, around 37,898 sentences of the Portuguese Wikipedia were processed with the grammars for corpora. All the processed sentences were introducing articles which, in the DBpedia [3] taxonomy, had one of the following types: *species*, *anatomical structure*, *chemical compound*, *disease*, *currency*, *drug*, *activity*, *language*, *music genre*, *colour*, *ethnic group* or *protein*. A pos-tagger was used in the extraction, but only to identify adjectives. Also, in an additional stage, we have used it to identify the grammatical categories of the arguments of the triples and we noticed that most of the relations extracted were between nouns. The evaluation of the extracted triples was performed by human judges, who classified samples with triples of each relation as correct or incorrect. The quantities of relations extracted, the proportion of correct triples, as well as the agreement values, are shown in Table 3.

Table 4. (Noun) thesaurus in numbers.

		TeP	OT	CLIP	TOP
Words	**Quantity**	17,158	5,819	23,741	30,554
	Ambiguous	5,867	442	12,196	13,294
	Most ambiguous	20	4	47	21
Synsets	**Quantity**	8,254	1,872	7,468	9,960
	Avg. size	3.51	3.37	12.57	6.6
	Biggest	21	14	103	277

To test the synset discovery procedure, we have made several experiments using the thesaurus of nouns TeP, OpenThesaurus.PT (OT), and also the noun synonymy instances of PAPEL which, after clustering, became the thesaurus CLIP. We also used TeP as the base thesaurus and merged it, first with OT, and then with CLIP, giving rise to the biggest noun thesaurus, TOP.

Table 4 has information on each one of the thesaurus, more precisely, the quantity of words, words belonging to more than one synset (ambiguous), the number of synsets where the most ambiguous word occurs, the quantity of synsets, the average synset size (number of words), and the size of the biggest synset. 519 synsets of CLIP and 480 of TOP were manually validated, each by two human judges who had to classify each synset as: correct, incorrect, or don't know[8]. Besides the average validation results and the agreement rates, Table 5 also contains the results considering only synsets of ten or less words, the less problematic (CLIP' and TOP').

Table 5. Results of manual synset validation.

	Sample	Correct	Incorrect	N/A	Agreement
CLIP	519 sets	65.8%	31.7%	2.5%	76.1%
CLIP'	310 sets	81.1%	16.9%	2.0%	84.2%
TOP	480 sets	83.2%	15.8%	1.0%	82.3%
TOP'	448 sets	86.8%	12.3%	0.9%	83.0%

Table 6. Results of triples mapping.

		Hypernym_of	Part_of	Member_of
Term-based triples		62,591	2,805	5,929
1st	**Mapped**	27,750	1,460	3,962
	Same synset	233	5	12
	Already present	3,970	40	167
Semi-mapped triples		7,952	262	357
2nd	**Mapped**	88	1	0
	Could be inferred	50	0	0
	Already present	13	0	0
Synset-based triples		23,572	1,416	3,783

The last set of results presented here regard using the first mapping procedure to map all the hypernym-of, part-of and member-of term-based triples of PAPEL to the synsets of TOP. Table 6 shows the mapping numbers. After the first phase, 33,172 triples had both of their terms assigned to a synset, and 10,530 had only one assigned. However, 4,427 were not really added, either because the same synset was assigned to both of the terms or because the triple had already been added after analysing other term-based triple. In the second phase, where hypernymy links were used, only 89 new triples were mapped and, from those, 13 had previously been added while other 50 triples were discarded or not attached because they could be inferred. Moreover, 19,638 triples were attached at least to a synset with only one term.

Table 7. Automatic validation of synset-based triples.

Relation	Sample size	Validation
Hypernymy_of	419 synsets	44,1%
Member_of	379 synsets	24,3%
Part_of	290 synsets	24,8%

The triples mapping was validated using Google web search engine to look for evidence on the synset-based triples. Once again, a set of natural language generic patterns, indicative of each relation, was defined. Then, for each triple $A\ R\ B$, each combination of terms $a \in A$ and $b \in B$ connected by a pattern indicative of R[9] was searched for. Table 7 shows the results obtained for each validated sample, according to the triple validation score, calculated by the following expression, where $found(A, B, R) = 1$ if evidence is found for the triple or 0 otherwise:

$$score = \frac{\sum_{i=1}^{|A|} \sum_{j=1}^{|B|} found(A, B, R)}{|A_i| * |B_j|}$$

The second mapping procedure was not evaluated yet, but, besides being more generic, it maps every triple and not only part of triple set.

6 CONCLUDING REMARKS

This ongoing research is an answer to the growing demand on semantically aware applications and addresses the lack of public domain lexico-semantic resources for Portuguese. The tools for knowledge extraction and the lexical ontology itself might be useful for

[8] In some context, all the words of a correct synset could have the same meaning, while for incorrect synsets, at least one word could never mean the same meaning as the others.

[9] Patterns used for part-of and member-of were the same because these relations can be expressed in very similar ways.

researchers and developers of applications in Portuguese. To extend the potential utilisation scenarios we are devising to export the ontology to several data representation formats, such as RDF/OWL models, because these are the W3C standard description languages for the Semantic Web. Furthermore, the later languages ease the browsing and visualisation of ontologies and have other useful features like reasoning capabilities.

Even though most of the methods presented here are not completely new, some of them have never targeted Portuguese. Nevertheless, we believe their application in the proposed sequence to be a suitable alternative to the manual creation of lexical resources in any language.

ACKNOWLEDGEMENTS

Hugo Gonçalo Oliveira is supported by FCT scholarship grant SFRH/BD/44955/2008. We would also like to thank Hernani Costa and the KDigg team for their participation in project discussions and in the development of the tools for corpora processing.

REFERENCES

[1] M. Berland and E. Charniak, 'Finding parts in very large corpora', in *Proc. 37th Annual Meeting of the ACL on Computational Linguistics*, pp. 57–64, Morristown, NJ, USA, (1999). Association for Computational Linguistics.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, 'The Semantic Web', *Scientific American*, (May 2001).

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, 'Dbpedia – a crystallization point for the web of data', *Web Semantics: Science, Services and Agents on the World Wide Web*, **7**(3), 154–165, (September 2009).

[4] C. Brewster and Y. Wilks, 'Ontologies, Taxonomies, Thesauri: Learning from Texts', in *Proc. The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop*, London, UK, (2004). Centre for Computing in the Humanities, Kings College.

[5] N. Calzolari, L. Pecchia, and A. Zampolli, 'Working on the italian machine dictionary: a semantic approach', in *Proc. 5th Conference on Computational Linguistics*, pp. 49–52, Morristown, NJ, USA, (1973). Association for Computational Linguistics.

[6] S. A. Caraballo, 'Automatic construction of a hypernym-labeled noun hierarchy from text', in *Proc. 37th annual meeting of the ACL on Computational Linguistics*, pp. 120–126, Morristown, NJ, USA, (1999). Association for Computational Linguistics.

[7] S. Cederberg and D. Widdows, 'Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction', in *Proc. 7th Conference on Computational Natural Language Learning (CoNLL)*, pp. 111–118, Morristown, NJ, USA, (2003). Association for Computational Linguistics.

[8] H. Costa, H. Gonçalo Oliveira, and P. Gomes, 'The impact of distributional metrics in the quality of relational triples', in *Proc. ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, (2010).

[9] G. de Melo and G. Weikum, 'On the utility of automatically generated wordnets', in *Proc. 4th Global WordNet Conference (GWC 2008)*, pp. 147–161, Szeged, Hungary, (2008). University of Szeged.

[10] P. C. de Oliveira, *Probabilistic Reasoning in the Semantic Web using Markov Logic*, Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, July 2009.

[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, **41**, 391–407, (1990).

[12] B. C. Dias-Da-Silva and H. R. de Moraes, 'A construção de um thesaurus eletrônico para o português do Brasil', *ALFA*, **47**(2), 101–115, (2003).

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

[14] B. Dorow, *A Graph Model for Words and their Meanings*, Ph.D. dissertation, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart, 2006.

[15] *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, ed., C. Fellbaum, The MIT Press, 1998.

[16] D. Gfeller, J.-C. Chappelier, and P. De Los Rios, 'Synonym Dictionary Improvement through Markov Clustering and Clustering Stability', in *Proc. of Intl. Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, pp. 106–113, (2005).

[17] R. Girju, A. Badulescu, and D. Moldovan, 'Automatic discovery of part-whole relations', *Computational Linguistics*, **32**(1), 83–135, (2006).

[18] R. Girju and D. Moldovan, 'Text mining for causal relations', in *Proc. 15th Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 360–364, (2002).

[19] H. Gonçalo Oliveira, D. Santos, and P. Gomes, 'Relations extracted from a portuguese dictionary: results and first evaluation', in *Local Proc. 14th Portuguese Conference on Artificial Intelligence (EPIA)*, (2009).

[20] M. A. Hearst, 'Automatic acquisition of hyponyms from large text corpora', in *Proc. 14th conference on Computational Linguistics*, pp. 539–545, Morristown, NJ, USA, (1992). Association for Computational Linguistics.

[21] A. Herbelot and A. Copestake, 'Acquiring ontological relationships from wikipedia using RMRS', in *Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*, (2006).

[22] G. Hirst, 'Ontology and the lexicon', in *Handbook on Ontologies*, eds., Steffen Staab and Rudi Studer, International Handbooks on Information Systems, 209–230, Springer, (2004).

[23] N. Ide and J. Veronis, 'Knowledge extraction from machine-readable dictionaries: An evaluation', in *Machine Translation and the Lexicon, LNAI*. Springer, (1995).

[24] A. Kilgarriff, '"I don't believe in word senses"', *Computing and the Humanities*, **31**(2), 91–113, (1997).

[25] D. Lin and P. Pantel, 'Concept discovery from text', in *Proc. Conference on Computational Linguistics (COLING)*, (2002).

[26] P. Marrafa, 'Portuguese Wordnet: general architecture and internal semantic relations', *DELTA*, **18**, 131–146, (2002).

[27] E. Navarro, F. Sajous, B. Gaume, L. Prévot, S. Hsieh, T. Y. Kuo, P. Magistry, and C. R. Huang, 'Wiktionary and nlp: Improving synonymy networks', in *Proc. Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Suntec, Singapore, (2009). Association for Computational Linguistics.

[28] R. Navigli, 'Word sense disambiguation: A survey', *ACM Computing Surveys*, **41**(2), (2009).

[29] E. Nichols, F. Bond, and D. Flickinger, 'Robust ontology acquisition from machine-readable dictionaries', in *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1111–1116. Professional Book Center, (2005).

[30] P. Pantel and M. Pennacchiotti, 'Automatically harvesting and ontologizing semantic relations', in *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, eds., Paul Buitelaar and Phillip Cimmiano, IOS Press, (2008).

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

[32] E. Riloff and J. Shepherd, 'A corpus-based approach for building semantic lexicons', in *Proc. 2nd Conference on Empirical Methods in Natural Language Processing*, pp. 117–124, (1997).

[33] M. Ruiz-Casado, E. Alfonseca, and P. Castells, 'Automatic assignment of wikipedia encyclopedic entries to wordnet synsets', in *Proc. Advances in Web Intelligence 3rd Intl. Atlantic Web Intelligence Conference (AWIC)*, pp. 380–386. Springer, (2005).

[34] A. Simões and R. Farinha, 'Dicionário Aberto: Um novo recurso para PLN', *Vice-Versa*, (April 2010). forthcomming.

[35] S. Soderland and B. Mandhani, 'Moving from textual relations to ontologized relations', in *Proc. AAAI Spring Symposium on Machine Reading*, (2007).

[36] P. D. Turney, 'Mining the web for synonyms: PMI–IR versus LSA on TOEFL', in *Proc. 12th European Conference on Machine Learning (ECML-2001)*, volume 2167, pp. 491–502. Springer, (2001).

[37] S. M. van Dongen, *Graph Clustering by Flow Simulation*, Ph.D. dissertation, University of Utrecht, 2000.

[38] P. Vossen, 'Eurowordnet: a multilingual database for information retrieval', in *Proc. DELOS workshop on Cross-Language Information Retrieval*, Zurich, (1997).

[39] T. W., E. Ovchinnikova, U. Krumnack, and H. Dittmann, 'Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology', in *Third Australasian Ontology Workshop (AOW 2007)*, volume 85 of *CRPIT*, pp. 61–69, Gold Coast, Australia, (2007). ACS.

[40] T. Zesch, C. Müller, and I. Gurevych, 'Extracting lexical semantic knowledge from Wikipedia and Wiktionary', in *Proc. 6th Intl. Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, (2008).