Towards the Automatic Creation of a Wordnet from a Term-based Lexical Network

Hugo Gonçalo Oliveira*
CISUC, University of Coimbra
Portugal
hroliv@dei.uc.pt

Paulo Gomes
CISUC, University of Coimbra
Portugal
pgomes@dei.uc.pt

Abstract

The work described here aims to create a wordnet automatically from a semantic network based on terms. So, a clustering procedure is ran over a synonymy network, in order to obtain synsets. Then, the term arguments of each relational triple are assigned to the latter, originating a wordnet. Experiments towards our goal are reported and their results validated.

1 Introduction

In order perform tasks where understanding the information conveyed by natural language is critical, today's applications demand better access to semantic knowledge. Knowledge about words and their meanings is typically structured in lexical ontologies, such as Princeton WordNet (Fellbaum, 1998), but this kind of resources is most of the times handcrafted, which implies much time-consuming human effort. So, the automatic construction of such resources arises as an alternative, providing less intensive labour, easier maintenance and allowing for higher coverage, as a trade-off for lower, but still acceptable, precision.

This paper is written in the scope of a project where several textual resources are being exploited for the construction of a lexical ontology for Portuguese. We have already made a first approach on the extraction of relational triples from text, where, likewise Hearst (1992), we take advantage of textual patterns indicating semantic relations. However, the extracted triples are held between two terms, which is not enough to build a lexical ontology capable of dealing with ambiguity.

Therefore, we present our current approach towards the automatic integration of lexico-semantic knowledge into a single independent lexical ontology, which will be structured on concepts and adopt a model close to WordNet's. The task of establishing synsets and mapping term-based triples to them is closely related to word sense disambiguation, where the only available context consists of the connections in the term-base network.

After contextualising this work, our approach is described. It involves (i) a clustering procedure for obtaining a thesaurus from a synonymy network, (ii) the augmentation of the later with manually created thesaurus, and (iii) mapping term-based relational triples to the thesaurus, to obtain a wordnet. Then, our experimentation results, as well as their validation, are presented. Briefly, we have tested the proposed approach on a term-based lexical network, extracted automatically from a dictionary. Synsets were validated manually while the attached triples were validated with the help of a web search engine.

2 Context

Our ultimate goal is the automatic construction of a broad-coverage structure of words according to their meanings, also known as a lexical ontology, the first subject of this section. We proceed with a brief overview on work concerned with moving from term-based knowledge to synset-based knowledge, often called ontologising.

2.1 Lexical Ontologies

Despite some terminological issues, lexical ontologies can be seen both as a lexicon and as an ontology (Hirst, 2004) and are significantly different from classic ontologies (Gruber, 1993). They are not based on a specific domain and are intended to provide knowledge structured on lexical items (words) of a language by relating them according to their meaning. Moreover, the main goal of a lexical ontology is to assemble lexical and semantic information, instead of storing common-sense knowledge (Wandmacher et al., 2007).

*supported by FCT scholarship SFRH/BD/44955/2008.

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 10–18,
Uppsala, Sweden, 16 July 2010.

Princeton WordNet (Fellbaum, 1998) is the most representative lexico-semantic resource for English and also the most accepted model of a lexical ontology. It is structured around groups of synonymous words (synsets), which describe concepts, and connections, denoting semantic relations between those groups. The success of WordNet led to the adoption of its model by lexical resources in different languages, such as the ones in the EuroWordNet project (Vossen, 1997), or WordNet.PT (Marrafa, 2002), for Portuguese.

However, the creation of a wordnet, as well as the creation of most ontologies, is typically manual and involves much human effort. Some authors (de Melo and Weikum, 2008) propose translating Princeton WordNet to wordnets in other languages, but if this might be suitable for several applications, a problem arises because different languages represent different socio-cultural realities, do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently (Hirst, 2004).

Another popular alternative is to extract lexico-semantic knowledge and learn lexical ontologies from text. Research on this field is not new and varied methods have been proposed to achieve different steps of this task including the extraction of semantic relations (e.g. (Hearst, 1992) (Girju et al., 2006)) or sets of similar words (e.g. (Lin and Pantel, 2002) (Turney, 2001)).

Whereas the aforementioned works are based on unstructured text, dictionaries started earlier (Calzolari et al., 1973) to be seen as an attractive target for the automatic acquisition of lexico-semantic knowledge. MindNet (Richardson et al., 1998) 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).

For Portuguese, PAPEL (Gonçalo Oliveira et al., 2009) is a lexical network consisting of triples denoting semantic relations between words found in a dictionary. Other Portuguese lexical ontologies, created by different means, are reviewed and compared in (Santos et al., 2009) and (Teixeira et al., 2010).

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 knowledge and have thus been receiving more and more attention by the community (see for instance (Zesch et al., 2008) (Navarro et al., 2009)).

2.2 Other Relevant Work

Most of the methods proposed to extract relations from text have term-based triples as output. Such a triple, *term1* RELATION *term2*, indicates that a possible meaning of *term1* is related to a possible meaning of *term2* by means of a RELATION.

Although it is possible to create a lexical network from the latter, this kind of networks is often impractical for computational applications, such as the ones that deal with inference. For instance, applying a simple transitive rule, a SYNONYM_OF $b \wedge b$ SYNONYM_OF $c \rightarrow a$ SYNONYM_OF c over a set of term-based triples can lead to serious inconsistencies. A curious example in Portuguese, where synonymy between two completely opposite words is inferred, is reported in (Gonçalo Oliveira et al., 2009): *queda* SYNONYM_OF *ruína* $\wedge$ *queda* SYNONYM_OF *habilidade* $\rightarrow$ *ruína* SYNONYM_OF *habilidade*. This happens because natural language is ambiguous, especially when dealing with broad-coverage knowledge. In the given example, *queda* can either mean *downfall* or *aptitude*, while *ruína* means *ruin*, *destruction*, *downfall*.

A possible way to deal with ambiguity is to adopt 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, bring together natural language and knowledge engineering in a suitable representation, for instance, for the Semantic Web (Berners-Lee et al., 2001). Of course that, from a linguistic point of view, word senses are complex and overlapping structures (Kilgarriff, 1997) (Hirst, 2004). 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.

In order to move from term-based triples to an ontology, Soderland and Mandhani (2007) describe a procedure where, besides other stages, terms in triples are assigned to WordNet synsets. Starting with all the synsets containing a term in

a triple, the term is assigned to the synset with higher similarity to the contexts from where the triple was extracted, computed based on the terms in the synset, sibling synsets and direct hyponym synsets.

Two other methods for ontologising term-based triples are presented by Pantel and Pennacchiotti (2008). One assumes that terms with the same relation to a fixed term are more plausible to describe the correct sense, so, to select the correct synset, it exploits triples of the same type sharing one argument. The other method, which seems to perform better, selects suitable synsets using generalisation through hypernymy links in WordNet.

There are other works where WordNet is enriched, for instance with information in its glosses, domain knowledge extracted from text (e.g. (Harabagiu and Moldovan, 2000) (Navigli et al., 2004)) or wikipedia entries (e.g. (Ruiz-Casado et al., 2005)), thus requiring a disambiguation phase where terms are assigned to synsets.

In the construction of a lexical ontology, synonymy plays an important role because it defines the conceptual base of the knowledge to be represented. One of the reasons for using WordNet synsets as a starting point for its representation is that, while it is quite straightforward to define a set of textual patterns indicative of several semantic relations between words (e.g. hyponymy, part-of, cause) with relatively good quality, the same does not apply for synonymy. In opposition to other kinds of relation, synonymous words, despite typically sharing similar neighbourhoods, may not co-occur frequently in unstructured text, especially in the same sentence (Dorow, 2006), leading to few indicative textual patterns. Therefore, most of the works on synonymy extraction from corpora rely on statistics or graph-based methods (e.g. (Lin and Pantel, 2002) (Turney, 2001) (Dorow, 2006)). Nevertheless, methods for synonymy identification based on co-occurrences (e.g. (Turney, 2001)) are more prone to identify similar words or near synonyms than real synonyms.

On the other hand, synonymy instances can be quite easily extracted from resources structured on words and meanings, such as dictionaries, by taking advantage not only of textual patterns, more frequent in those resources (e.g. *também conhecido por/como*, *o mesmo que*, for Portuguese), but also of definitions consisting of only one word or a enumeration, which typically contain synonyms of the defined word. So, as it is possible to create a lexical network from a set of relational triples ($a\ R\ b$), a synonymy network can be created out of synonymy instances (a SYNONYM_OF b). Since these networks tend to have a clustered structure, Gfeller et al. (2005) propose a clustering procedure to improve their utility.

3 Research Goals

The research presented here is in the scope of a project whose final goal is to create a lexical ontology for Portuguese by automatic means. Although there are clear advantages of using resources already structured on words and meanings, dictionaries are static resources which contain limited knowledge and are not always available for this kind of research. On the other hand, there is much text available on the most different subjects, but free text has few boundaries, leading to more ambiguity and parsing issues.

Therefore, it seems natural to create a lexical ontology with knowledge from several textual sources, from (i) high precision structured resources, such as manually created thesaurus, to (ii) semi-structured resources such as dictionaries or collaborative encyclopedias, as well as (iii) unstructured textual corpora. Likewise Wandmacher et al. (2007) propose for creating a lexical ontology for German, these are the general lines we will follow in our research, but for Portuguese.

Considering each resource specificities, including its organisation or the vocabulary used, the extraction procedures might be significantly different, but they should all have one common output: a set of term-based relational triples that will be integrated in a single lexical ontology.

Whereas the lexical network established by the triples could be used, these networks are not suitable for several tasks, as discussed in Section 2.2. A fragment of a synonymy network extracted from a Portuguese dictionary can be seen in Figure 1. Since all the connections imply synonymy, the network suggests that all the words are synonymous, which is not true. For example, the word *copista* may have two very distinct meanings: (a) a person who writes copies of written documents or (b) someone who drinks a lot of wine. On the other hand, other words which may refer to the same concept as, for instance, meaning (a) of *copista*, such as *escrevente*, *escrivão* or *transcritor*.

So, in order to deal with ambiguity in natural

language, we will adopt a wordnet-like structure which enables the establishment of unambiguous semantic relations between synsets.

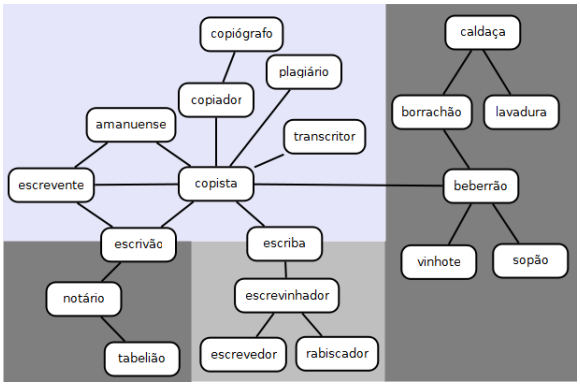

Figure 1: Fragment of a synonymy network.

4 Approach

Considering our goal, a set of term-based triples goes through the following stages: (i) clustering over the synonymy network for the establishment of synsets, to obtain a thesaurus; (ii) augmentation of the thesaurus by merging it with synsets from other resources; (iii) assignment of each argument of a term-based triple (except synonymy) to a synset in the thesaurus, to obtain a wordnet. Note that stages (i) and (ii) are not both mandatory, but at least one must be performed to obtain the synsets.

Looking at some of the works referred in Section 2.2, ours is different because it does not require a conceptual base such as WordNet. Also, it integrates knowledge from different sources and tries to disambiguate each word using only knowledge already extracted and not the context where the word occurs.

4.1 Clustering for a thesaurus

This stage was originally defined after looking at disconnected pieces of a synonymy network extracted from a dictionary, which had a clustered structure apparently suitable for identifying synsets. This is also noticed by Gfeller et al. (2005) who have used the Markov Clustering algorithm (MCL) (van Dongen, 2000) to find clusters in a synonymy network.

Therefore, since MCL had already been applied to problems very close to ours (e.g. (Gfeller et al., 2005), (Dorow, 2006)), it seemed to suit our purpose – it would not only organise a term-based network into a thesaurus, but, if a network extracted from several resources is used, clustering would homogenise the synonymy representation.

MCL finds clusters by simulating random walks within a graph by alternately computing random walks of higher length, and increasing the probabilities of intra-cluster walks. It can be briefly described in five steps: (i) take the adjacency matrix A of the graph; (ii) normalise each column of A to 1 in order to obtain a stochastic matrix S; (iii) compute S^2; (iv) take the γth power of every element of S^2 and normalise each column to 1[1]; (v) go back to (ii) util MCL converges to a matrix idempotent under steps (ii) and (iii).

Since MCL is a hard-clustering algorithm, it assigns each term to only one cluster thus removing ambiguities. To deal with this, Gfeller et al. (2005) propose an extension to MCL for finding unstable nodes in the graph, which frequently denote ambiguous words. This is done by adding random stochastic noise, δ, to the non-zero entries of the adjacency matrix and then running MCL with noise several times. Looking at the clusters obtained by each run, a new matrix can be filled based on the probability of each pair of words belonging to the same cluster.

We have adopted this procedure, with slight differences. First, we observed that, for the network we used, the obtained clusters were closer to the desired results if $-0.5 < \delta < 0.5$. Additionally, in the first step of MCL, we use frequency-weighted adjacency matrixes F, where each element F_{ij} corresponds to the number of existing synonymy instances between i and j. Although using only one dictionary each synonymy instance will be extracted at most two times (a SYNONYM_OF b and b SYNONYM_OF a), if more resources are used, it will strengthen the probability that two words appearing frequently as synonyms belong to the same cluster.

Therefore, the clustering stage has the following steps: (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, with $\gamma = 1.6$, 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 probabil-

[1]Increasing γ (typically $1.5 < \gamma < 2$) increases the granularity of the clusters.

ities of each pair of words in F belonging to the same cluster; (v) create the clusters based on P and on a given threshold $\theta = 0.2$. If $P_{ij} > \theta$, 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. Applying this procedure to the network in Figure 1 results in the four represented clusters. There, ambiguous words *escrivão* and *escriba* are included in two different clusters.

4.2 Merging synsets for thesaurus augmentation

In this stage, other resources with synsets, such as manually created thesaurus, are merged together and then merged with the thesaurus obtained in the previous stage, by the following procedure: (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[2] 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 $|J| = 1$, the synset in J is copied 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 created, and all elements of J are removed from M.

The synsets of the resulting thesaurus will be used as the conceptual base in which the term-based triples are going to be mapped.

4.3 Assigning terms to synsets

After the previous stages, the following are available: (i) a thesaurus T and (ii) a term-based semantic network, N, where each edge has a type, R, and denotes a semantic relation held between the meaning of the terms in the two nodes it connects. Using T and N, this stage tries to map term-based triples to synset-based triples, or, in other words, assign each term, a and b, in each triple, $(a\ R\ b) \in N$, to suitable synsets. The result is a knowledge base organised as a wordnet.

In order to assign a to a synset A, b is fixed and all the synsets containing a, $S_a \subset T$, are collected. If a is not in the thesaurus, 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[3]. Then, $p_{ai} = \frac{n_{ai}}{|S_{ai}|}$ is calculated. Finally, all the synsets with the highest p_{ai} are added to C and (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 $p_{ai} < \theta$ [4]; (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.

If hypernymy links are already established, semi-mapped triples, where one of the arguments is assigned to a synset and the other is not, $(A\ R\ b)$ or $(a\ R\ B)$, go to a second phase. There, hypernymy is exploited together with the assignment candidates, in C, to help assigning the unassigned term in each semi-mapped triple, or to remove triples that can be inferred. Take for instance $(A\ R\ b)$. If there is one synset $C_i \in C$ with:

- a hypernym synset H, (H HYPERNYM_OF C_i) and a triple $(A\ R\ H)$, b would be assigned to C_i, but, since hyponyms inherit all the properties of their hypernym, the resulting triple can be inferred and is thus ignored: $(A\ R\ H) \wedge (H \text{ HYPERNYM\_OF } C_i) \rightarrow (A\ R\ C_i)$[5]

  For example, if H=(*mammal*) and C_i=(*dog*), possible values of A and R are A=(*hair*) R=PART_OF; A=(*animal*) R=HYPERNYM_OF

- a hyponym synset H, (C_i HYPERNYM_OF H) and a triple $(A\ R\ H)$, b is assigned to C_i. Furthermore, if all the hyponyms of C_i, (C_i HYPERNYM_OF I_i), are also related to A in the same way, $(A\ R\ I_i)$, it can be inferred that I_i inherits the relation from C_i. So, all the later triples can be inferred and thus removed.

  For example, if H=(*dog*), I_i=(*cat*), I_j=(*mouse*) and C_i=(*mammal*), possible values of A and R are A=(*hair*) R=PART_OF; A=(*animal*) R=HYPERNYM_OF

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

[3] If R is a transitive relation, the procedure may benefit from applying one level of transitivity to the network: $x\ R\ y \wedge y\ R\ z \rightarrow x\ R\ z$. However, since relations are held between terms, some obtained triples might be incorrect. So, although the latter can be used to help selecting a suitable synset, they should not be mapped to synsets themselves.

[4] 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

[5] Before applying these rules it is necessary to make sure that all relations are represented only in one way, otherwise they might not work. For instance, if the decision is to represent *part-of* triples in the form $part$ PART_OF $whole$, triples $whole$ HAS_PART $part$ must be reversed. Furthermore, these rules assume that hypernymy relations are all represented $hypernym$ HYPERNYM_OF $hyponym$ and not $hyponym$ HYPONYM_OF $hypernym$.

5 Experimentation

In this section we report experimental results obtained after applying our procedure to part of the lexical network of PAPEL (Gonçalo Oliveira et al., 2009). The clustering procedure was first ran over PAPEL's noun synonymy network in order to obtain the synsets which were later merged with two manually created thesaurus. Finally, hypernym-of, member-of and part-of triples of PAPEL were mapped to the thesaurus by assigning a synset to each term argument.

5.1 Resources used

For experimentation purposes, freely available lexical resources for Portuguese were used. First, the last version of PAPEL, 2.0, a lexical network for Portuguese created automatically from a dictionary, as referred in Section 2. PAPEL 2.0 contains approximately 100,000 words, identified by their orthographical form, and approximately 200,000 term-based triples relating the words by different types of semantic relations.

In order to enrich the thesaurus obtained from PAPEL, TeP (Dias-Da-Silva and de Moraes, 2003) and OpenThesaurus.PT[6] (OT), were used. Both of them are manually created thesaurus, for Brazilian Portuguese and European Portuguese respectively, modelled after Princeton WordNet (Fellbaum, 1998) and thus containing synsets. Besides being the only freely available thesaurus for Portuguese we know about, TeP and OT were used together with PAPEL because, despite representing the same kind of knowledge, they are mostly complementary, which is also observed by (Teixeira et al., 2010) and (Santos et al., 2009).

Note that, for experimentation purposes, we have only used the parts of these resources concerning nouns.

5.2 Thesaurus creation

The first step for applying the clustering procedure is to create PAPEL's synonymy network, which is established by its synonymy instances, a SYNONYM_OF b. After splitting the network into independent disconnected sub-networks, we noticed that it was composed by a huge sub-network, with more than 16,000 nodes, and several very small networks. If ambiguity was not resolved, this would suggest that all the 16,000 words had the same meaning, which is not true.

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

		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

Table 1: (Noun) thesaurus in numbers.

		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

Table 2: Results of triples mapping

A small sample of this problem can be observed in Figure 1.

We then ran the clustering procedure and the thesaurus of PAPEL, CLIP, was obtained. Finally, we used TeP as the base thesaurus and merged it, first with OT, and then with CLIP, giving rise to the noun thesaurus we used in the rest of the experimentation, TOP.

Table 1 contains information about 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[7].

5.3 Mapping the triples

The mapping procedure was applied to all the hypernym-of, part-of and member-of term-based triples of PAPEL, distributed according to Table 2 where additional numbers on the mapping are presented. After the first phase of the mapping, 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, 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. Another interesting fact is that 19,638 triples were attached to a synset with only one term. From those, 5,703 had a synset

[7] Synsets with only one word were ignored in the construction of Table 1.

with only one term in both arguments.

We ended up with a wordnet with 27,637 synsets, 23,572 hypernym-of, 1,416 part-of and 3,783 member-of synset-based triples.

6 Validation of the results

Evaluation of a new broad-coverage ontology is most of the times performed by one of two means: (i) manual evaluation of a representative subset of the results; (ii) automatic comparison with a gold standard. However, while for English most researchers use Princeton WordNet as a gold standard, for other languages it is difficult to find suitable and freely available consensual resources. Considering Portuguese, as we have said earlier, TeP and OT are effectively two manually created thesaurus but, since they are more complementary than overlapping to PAPEL, we thought it would be better to use them to enrich our resource.

There is actually a report (Raman and Bhattacharyya, 2008) with an automatic evaluation of synsets, but we decided no to follow it because this evaluation is heavily based on a dictionary and we do not have unrestricted access to a full and updated dictionary of Portuguese and also, indirectly by PAPEL, a dictionary was one of our main sources of information.

Therefore, our choice relied on manual validation of the synsets of CLIP and TOP. Furthermore, synset-based triples were validated in an alternative automatic way using a web search engine.

6.1 Manual validation of synsets

Ten reviewers took part in the validation of ten random samples with approximately 50 synsets from each thesaurus. We made sure that each synset was not in more than one sample and synsets with more than 50 terms were not validated. Also, in order to measure the reviewer agreement, each sample was analysed by two different reviewers. Given a sample, each reviewer had to classify each synset as: correct (1), if, in some context, all the terms of the synset could have the same meaning, or incorrect (0), if at least one term of the synset could never mean the same as the others. The reviewers were advised to look for the possible meanings of each word in different dictionaries. Still, if they could not find them, or if they did not know how to classify the synset, they had a third option, N/A (2).

In the end, 519 synsets of CLIP and 480 of TOP were validated. When organising the validation results we noticed that the biggest synsets were the ones with more problems. So, besides the complete validation results, Table 3 also contains the results considering only synsets of ten or less words, when a ' is after the name of the thesaurus. The presented numbers are the average between the classifications given by the two reviewers and the agreement rate corresponds to the number of times both reviewers agreed on the classification.

Even though these results might be subjective, since they are based on the reviewers criteria and on the dictionaries they used, they can give an insight on the quality of the synsets. The precision results are acceptable and are improved if the automatically created thesaurus is merged with the ones created manually, and also when bigger synsets are ignored. Most of the times, big synsets are confusing because they bring together more than one concept that share at least one term. For instance, take the synset: *insobriedade, desmedida, imoderação, excesso, nimiedade, desmando, desbragamento, troco, descontrolo, superabundância, desbunda, desregramento, demasia, incontinência, imodicidade, superação, intemperança, descomedimento, superfluidade, sobejidão, acrasia*, where there is a mix of the concepts: (a) insobriety, not following all the rules, heedless of the consequences and, (b) surplus. Both of these concepts can be referred to as an *excess* (*excesso*).

6.2 Automatic validation of triples

The automatic validation of the triples attached to our wordnet consisted of using Google web search engine to look for evidence on their truth. This procedure started by removing terms whose occurrences in Google were less than 5,000. Synsets that became empty were not considered and, from the rest, a sample was selected for each one of the three types of relation.

Following the idea in (Gonçalo Oliveira et al., 2009), a set of natural language generic patterns, indicative of each relation, was defined having in mind their input to Google[8]. Then, for each triple (A R B), the patterns were used to search for ev-

[8]Hypernymy patterns included: [hypo] *é um|uma (tipo|forma|variedade|...)* de* [hyper], [hypo] *e outros|outras* [hyper] or [hyper] *tais como* [hypo]. Patterns for part-of and member-of were the same because these relations can be expressed in very similar ways, and included: [part/member] *é (parte|membro|porção) do|da* [whole/group], [part/member] *(faz parte)* do|da* [whole/group] or [whole/group] *é um (grupo|conjunto|...) de* [part/member].

	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 3: Results of manual synset validation.

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

Table 4: Automatic validation of triples

idence on each combination of terms $a \in A$ and $b \in B$ connected by a pattern indicative of R. The triple validation score was then calculated by expression 1, where $found(A, B, R) = 1$ if evidence is found for the triple or 0 otherwise.

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

Table 4 shows the results obtained for each validated sample. Pantel and Pennacchiotti (2008) perform a similar task and present precision results for part-of (40.7%-57.4%) and causation (40.0%-45%) relations. It is however not possible to make a straight comparison. For their experimentation, they selected only correct term-based triples extracted from text and their results were manually validated by human judges. On the other hand, we have used term-based triples extracted automatically from a dictionary, with high but not 100% precision, from where we did not choose only the correct ones, and we have used synsets obtained from our clustering procedure which, once again, have lower precision. Moreover, we validated our results with Google where, despite its huge dimension, there are plenty of ways to denote a semantic relation, when we had just a small set textual patterns. Also, despite occurring more than 5,000 times in Google, some terms correctly included in a synset were conveying less common meanings.

Nevertheless, we could not agree more with Pantel and Pennacchiotti (2008) who state that attaching term-based triples to an ontology is not an easy task. Therefore, we believe our results to be promising and, if more refined rules are added to our set, which is still very simple, they will surely be improved.

7 Concluding remarks

We have presented our first approach on two crucial steps on the automatic creation of a wordnet lexical ontology. Clustering proved to be a good alternative to create a thesaurus from a dictionary's synonymy network, while a few rules can be defined to attach a substantial number of term-based triples to a synset based resource.

Despite interesting results, in the future we will work on refining the attachment rules and start integrating other relations such as causation or purpose. Furthermore, we are devising new methods for attaching terms to synsets. For instance, we have recently started to do some experiences with an attaching method which uses the lexical network's adjacency matrix to find the most similar pair of synsets, each of them containing one of the arguments of a term-based triple.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American*, May.

Nicoletta Calzolari, Laura Pecchia, and Antonio Zampolli. 1973. Working on the italian machine dictionary: a semantic approach. In *Proc. 5th Conference on Computational Linguistics*, pages 49–52, Morristown, NJ, USA. Association for Computational Linguistics.

Gerard de Melo and Gerhard Weikum. 2008. On the utility of automatically generated wordnets. In *Proc. 4th Global WordNet Conf. (GWC)*, pages 147–161, Szeged, Hungary. University of Szeged.

Bento Carlos Dias-Da-Silva and Helio Roberto de Moraes. 2003. A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

Beate Dorow. 2006. *A Graph Model for Words and their Meanings*. Ph.D. thesis, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

David Gfeller, Jean-Cédric Chappelier, and Paulo De Los Rios. 2005. Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In *Proc. of International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, pages 106–113.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Hugo Gonçalo Oliveira, Diana Santos, and Paulo Gomes. 2009. Relations extracted from a portuguese dictionary: results and first evaluation. In *Local Proc. 14th Portuguese Conf. on Artificial Intelligence (EPIA)*.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Sanda M. Harabagiu and Dan I. Moldovan. 2000. Enriching the wordnet taxonomy with contextual knowledge acquired from text. In *Natural language processing and knowledge representation: language for knowledge and knowledge for language*, pages 301–333. MIT Press, Cambridge, MA, USA.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th Conf. on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Graeme Hirst. 2004. Ontology and the lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer.

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

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proc. 19th Intl. Conf. on Computational Linguistics (COLING)*, pages 577–583.

Palmira Marrafa. 2002. Portuguese Wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Tzu Y. Kuo, Pierre Magistry, and Chu R. Huang. 2009. Wiktionary and nlp: Improving synonymy networks. In *Proc. Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, Suntec, Singapore. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, Alessandro Cucchiarelli, and Francesca Neri. 2004. Extending and enriching wordnet with ontolearn. In *Proc. 2nd Global WordNet Conf. (GWC)*, pages 279–284, Brno, Czech Republic. Masaryk University.

Patrick Pantel and Marco Pennacchiotti. 2008. Automatically harvesting and ontologizing semantic relations. In Paul Buitelaar and Phillip Cimmiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press.

J. Raman and Pushpak Bhattacharyya. 2008. Towards automatic evaluation of wordnet synsets. In *Proc. 4th Global WordNet Conf. (GWC)*, pages 360–374, Szeged, Hungary. University of Szeged.

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. In *Proc. 17th Intl. Conf. on Computational Linguistics (COLING)*, pages 1098–1102.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *Proc. Advances in Web Intelligence Third Intl. Atlantic Web Intelligence Conf. (AWIC)*, pages 380–386. Springer.

Diana Santos, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros, and Rosário Silva. 2009. O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL. In *Actas do XXV Encontro Nacional da Associação Portuguesa de Linguística (APL)*. forthcomming.

Stephen Soderland and Bhushan Mandhani. 2007. Moving from textual relations to ontologized relations. In *Proc. AAAI Spring Symposium on Machine Reading*.

Jorge Teixeira, Luís Sarmento, and Eugénio C. Oliveira. 2010. Comparing verb synonym resources for portuguese. In *Computational Processing of the Portuguese Language, 9th Intl. Conference, Proc. (PROPOR)*, pages 100–109.

Peter D. Turney. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *Proc. 12th European Conf. on Machine Learning (ECML)*, volume 2167, pages 491–502. Springer.

S. M. van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Piek Vossen. 1997. Eurowordnet: a multilingual database for information retrieval. In *Proc. DELOS workshop on Cross-Language Information Retrieval*, Zurich.

Tonio Wandmacher, Ekaterina Ovchinnikova, Ulf Krumnack, and Henrik Dittmann. 2007. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In *Third Australasian Ontology Workshop (AOW)*, volume 85 of *CRPIT*, pages 61–69, Gold Coast, Australia. ACS.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proc. 6th Intl. Language Resources and Evaluation (LREC)*, Marrakech, Morocco.