 ISBN: 978-989-95618-4-7

Ontologising Relational Triples into a Portuguese Thesaurus

Hugo Gonçalo Oliveira, Paulo Gomes
{hroliv,pgomes}@dei.uc.pt

CISUC, University of Coimbra, Portugal

Abstract. Having in mind the automatic acquisition and integration of knowledge from different heterogeneous resources, this paper proposes several automatic methods for attaching term-based relational triples to the synsets of a thesaurus, without exploiting the extraction context for disambiguation. After using the proposed methods to attach triples, extracted from a Portuguese dictionary, to the synsets of a Portuguese thesaurus, two perspectives on their performance are given. The resulting synset-based triples were automatically validated based on support found in corpus and were then evaluated based on a handcrafted gold resource.

Keywords: semantic relations, synsets, relational triples, lexical ontologies, ontologising

1 Introduction

Today's information extraction (IE) systems are capable of acquiring concepts and information about them from large collections of text. Whether these systems aim for the automatic acquisition of lexico-semantic relations (e.g. [11] [18]), knowledge on specific domains, or the extraction of open-domain facts (e.g. [1] [4]) they typically represent concepts as terms. A relational triple (a R b) is a common way of denoting a semantic relation where the arguments (a and b) are terms whose meaning is connected by a relation described by R. We will refer to the former as term-based triples (tb-triples). However, a simple term is usually not enough to unambiguously refer to a concept because the same word might have different meanings and different words might have the same meaning.

On the one hand, this problem is not severe in the extraction of domain knowledge, where, based on the one sense per discourse assumption [6], ambiguity is low. On the other hand, when dealing with broad-coverage knowledge, if ambiguities are not handled, it becomes impractical to formalise the extracted information and to accomplish tasks such as discovering new knowledge.

Therefore, to make IE systems more useful, a new step, which can be seen as a kind of word-sense disambiguation (WSD), is needed. Usually referred to as ontologising [17], this step aims at moving towards an ontological structure by associating the extracted terms to their meanings.

However, whereas most WSD techniques rely on the context where the words to be disambiguated occur to find their most adequate sense, we aim to achieve

WSD without having to represent this context. One of the main reasons behind this view is the possibility of having two independent modules in a IE system: one responsible for extracting tb-triples and other for ontologising the latter. In other words, the second module attaches each term in a triple to a concept, represented, for instance, as a synset in a broad-coverage lexical ontology. We believe that this approach is an interesting way of coping with information sparsity, since it will allow for the extraction of knowledge from different heterogeneous sources (e.g. dictionaries, encyclopedias, corpora), and provide a way to harmoniously integrate all the extracted information in a common knowledge base.

In this paper, we propose two methods for ontologising tb-triples and compare them with two baselines. We have used them for Portuguese, where there is a lack of broad-coverage knowledge bases, especially those providing semantic relations and glosses describing concepts. It is thus harder to represent the context of the available concepts. In order to see how far this approach could go, an experiment was carried out using the methods for ontologising tb-triples, automatically extracted from a Portuguese dictionary, in a synset base created after merging two Portuguese thesauri. The results obtained were automatically validated according to support found in a corpus and compared to a handcrafted gold resource, created especially for this evaluation.

In the rest of the paper, after presenting some related work, we introduce the ontologising methods. Then, we describe our experimentation. Before concluding, we present two perspectives on the evaluation of the methods and discuss on their performance, with special focus on their comparison.

2 Related Work

Word Sense Disambiguation (WSD) is the task of selecting the most adequate sense of a word in a particular context, usually from a list of possible senses, as in a dictionary or in a lexical ontology [15]. For English, Princeton WordNet [5] is usually used as a sense inventory (e.g. [21]). The goal of the disambiguation is thus to assign one of the possible WordNet senses of a word to an occurrence of this word in a context. Furthermore, the WordNet knowledge base can be used as an external source of knowledge for the achievement of WSD (e.g. [2]).

The research presented here is more specific and does not require a full disambiguation of text. It aims to move from knowledge structures based on terms, identified by their orthographical form, to ontological structures, which handle ambiguities and are thus aware of word meanings. This task was originally baptised as ontologising when a method was presented to link terms, extracted from text, to WordNet concepts, after inducing several features of their senses [17].

Besides enabling the formalisation of knowledge according to words and their meanings, providing tasks such as inference, these methods can be used to enrich ontologies with information automatically extracted from text. For instance, WordNet has been extended with domain knowledge, Wikipedia entries and also with information extracted from its own sense glosses (see [16] [20] [22] [10]). The new terms are associated with the synset(s) to which their context shares most

similarities. The context of each synset is represented, for instance, by the words in its glosses, the (synonymous) words it contains and by words in sibling synsets or synsets connected by hypernymy or hyponymy. When it comes to representing the context of a Wikipedia entry, besides the words used to describe its subject, it is possible to use the categories of the entry or outgoing links.

In a work similar to ours [25], stages for moving from term-based triples to triples, established between WordNet synsets are described. After normalising the arguments of a triple, each of its term arguments is associated to a synset that contains this term and has the most similar context to the triple. The latter context is represented by the words in the sentences from where the triple was extracted, while the context of the synset is represented in a similar fashion to the aforementioned works, namely using the words in the synset, in sibling synsets and in direct hyponyms.

In an even closer work [19], two methods to ontologise term-based triples taking advantage of the extracted information and of the structure of WordNet are proposed. The anchor approach assumes that words related in the same way to a fixed word are more plausible to describe the same sense – to select the correct synset, it exploits extracted triples of the same type sharing one term argument. The clustering approach selects suitable synsets using generalisation through hypernymy links in WordNet.

3 Ontologising Methods

In this section, we propose several methods for ontologising term-based triples (tb-triples), $w_a\ R\ w_b$, in a thesaurus T. In other words, the purpose of these methods is to attach terms w_a and w_b to suitable synsets $A_i \in T$ and $B_j \in T$, resulting in a synset-based triple (hereafter sb-triple), $A_i\ R\ B_j$. Having in mind the goal of this research, the proposed methods do not consider the context where the triples were extracted from, nor the glosses of the synsets[1].

However, in order to work properly, two of the methods take advantage of all the tb-triples given as intput. More precisely, these methods use the information in a given lexical network to select the best candidate synsets. A lexical network is established by a set of tb-triples, and is defined as a graph structure, $G = (N, E)$, with $|N|$ nodes and $|E|$ edges, $E \subset N^2$, where each node $w_i \in N$ represents a word and each edge between nodes w_i and w_j, $E(w_i, w_j)$, indicates that one of the meanings of the word w_i is related to one of the meanings of w_j. Furthermore, the edges can be labelled according to the type of relationship held by the two words, which can thus be obtained from a labelled tb-triple, such as $w_a\ R\ w_b$, where w_a and w_b are the words in the nodes and R describes the type of relation. By default, when a lexical network is needed, it is created from the triples given as input. These methods are thus better suited to ontologise large amounts of knowledge at the same time. So, when there are few input tb-triples, it is possible to provide a larger lexical network. The latter can be obtained, for instance, from

[1] As far as we know, there are no free thesaurus of Portuguese with available glosses.

an external resource or, eventually, from the ontology where the triples are being attached to, if the former contains already ontologised triples.

Even though there is not such a thesaurus, all the proposed methods assume that the thesaurus T is broad enough to cover all the senses of the words it contains. Therefore, in all of them, if a term refers to a monosemous word, this term is attached to the only possible sense of this word. Still, if T does not contain the argument of a tb-triple, a new synset containing only this term is created. As for the rest of the triples, each method has a different strategy, which starts by getting all the synsets containing term w_a, $A : \forall(A_i \in A) w_a \in A_i$, and all synsets with term w_b, $B : \forall(B_j \in B) w_b \in B_j$. We now present the methods:

Random : This method is used as a baseline and attaches term w_a to a random synset of A and term w_b to a random synset of B.

Average Frequency (AF) : A typical baseline in WSD relies on choosing the most frequent sense of a word [7]. However, while in Princeton WordNet synsets are ranked according to the most used senses, we do not have this information for Portuguese. This method is also based on frequencies but not on word senses. It attaches terms w_a and w_b to the highest ranked synset according to the average number of occurrences of its terms in a web search engine.

Related Proportion (RP) : This method is based on a similar assumption to the anchor approach [19] (see section 2). First, to attach term w_a to a synset, term w_b is fixed. Second, for each synset $A_i \in A$, n_i is the number of terms $w_k \in A_i$ such that $w_k \; R \; w_b$ holds. Then, the related proportion $p_i = \frac{n_i}{|A_i|}$ is calculated. All the candidate synsets with the highest p_i are added to set C. Finally,

- if $|C| = 1$, w_a is attached to the only synset in C, C_1;
- if $|C| > 1$, C' is the set of synsets in C with the highest n_i. If $|C'| = 1$, w_a is assigned to C'_1, unless $p_i < \theta$, a threshold defined to avoid that w_a is assigned to a big synset where w_a, itself, is the only term related to w_b;
- if it is not possible to attach w_a to a synset, it remains unassigned.

Term w_b is attached to a synset using the same procedure, but fixing w_a.

Average Cosine (AC) : This method also assumes that concepts in a semantic relation are described by words related with the same concepts. However, besides relations of the same type of the triple to be ontologised, it exploits other kinds of relations to identify the most similar synsets and select the best pair.

A term-term matrix M is first created based on the adjacencies of the given lexical network. Consequently, M is a square matrix with n lines, where n is the total number of nodes (terms) in the network. If the terms w_i and w_j are connected by one of the relations considered, $M_{ij} = 1$, otherwise, $M_{ij} = 0$.

In order to ontologise w_a and w_b, this method selects the most similar pair of synsets, $A_i \in A$ and $B_j \in B$, according to the adjacencies of the terms they contain. Therefore, the similarity between A_i and B_j, represented by the adjacency vectors of their terms, $M_{Ai} = (M_{Ai0}, ..., M_{Ain}), n = |A_i|$ and

$M_{Bj} = (M_{Bj0}, ..., M_{Bjn}), n = |B_j|$, is given by the average lexical network based similarity for each term $A_{ik} \in A_i$ with each term $B_{jl} \in B_j$:

$$sim(A_i, B_j) = \frac{\sum_{k=1}^{|A_i|} \sum_{l=1}^{|B_j|} \cos(M_{Aik}, M_{Bjl})}{|A_i||B_j|}$$

While this expression has been used to find similar nouns, represented as co-occurrence vectors in a corpus [3], we adapted it to measure the similarity of two synsets, in our case represented as the adjacency vectors of their terms.

4 Experimentation

In order to analyse the methods introduced in section 3 and to compare their results, we have selected samples of random relational triples automatically extracted from a Portuguese dictionary. All the methods were used to ontologise the triples into a Portuguese broad-coverage wordnet-like thesaurus, where concepts are represented as synsets. This section describes the resources used in this experimentation and reports on the coverage of each method.

4.1 Synsets

Currently, for Portuguese, there are two freely available broad-coverage synset-based thesauri, both created manually: TeP [13], an electronic thesaurus for Brazilian Portuguese, and OpenThesaurus (OT), a thesaurus for OpenOffice. The current version of TeP, 2.0[2], contains about 17,100 unique nouns, organised in about 8,200 synsets, while OT[3] contains about 6,100 nouns organised in about 2,000 synsets. Despite TeP being much larger, we decided to merge it with OT because TeP was made for Brazilian Portuguese and contains some unusual words/meanings in European Portuguese. Also, even though it is smaller, OT contains several words/meanings which are not covered by TeP.

As TeP is larger, we have used it as a starting point to automatically create a new thesaurus, TePOT. For each synset in OT, the most similar synset in TeP was selected, based on the following similarity measures, where A and B represent synsets:

$$Overlap = \frac{A \cap B}{min(|A|, |B|)} \qquad Jaccard(A, B) = \frac{A \cap B}{A \cup B}$$

The Overlap measure is used to select the first set of candidates. If the latter set is empty, the OT synset is copied to TePOT. Otherwise, it selects the candidate(s) with higher Jaccard coefficient and merges them first in one synset, and then with the OT synset. In the end, TePOT contains 18,501 nouns,

[2] Available through http://www.nilc.icmc.usp.br/tep2/
[3] Available through http://openthesaurus.caixamagica.pt/

organised in 8,293 synsets – 6,237 of the nouns are ambiguous and, in average, one synset has 3.84 terms and one term is in 1.72 synsets.

Furthermore, as referred in section 3, the synsets of TePOT were ranked according to their frequency. This value was computed by the average number of hits of its terms in Google web search engine. The frequency $Freq(S)$ of synset $S = (w_1, w_2, ...w_n)$ is thus given by the following expression:

$$Freq(S) = \frac{\sum_{i=1}^{|S|} Hits(w_i)}{|S|}$$

4.2 Term-based triples

The tb-triples used were obtained from the most recent version of PAPEL [9], 2.0[4] , which is a lexical network extracted automatically from a Portuguese dictionary, based on the exploitation of several systematic lexical patterns used in the definitions. PAPEL was created automatically, so its precision is not 100%, but it is probably the largest freely available source of structured lexico-semantic knowledge of Portuguese. Still, in order to minimise the noise, we only used triples supported by CETEMPúblico [24], a newspaper corpus of Portuguese. This means that the arguments of these triples co-occurred at least once in the corpus, connected by discriminating textual patterns for their relation. Furthermore, we discarded triples with very frequent and abstract arguments, such as *acto* (act), *efeito* (effect), *acção* (action), *estado* (state) and *coisa* (thing), as well as triples with arguments with less than 25 occurrences in CETEMPúblico.

This resulted in four samples of different semantic relations established between nouns: 500 hypernymy triples, 199 part-of triples, 436 member-of triples and 125 purpose-of triples. Together with the samples, parts of PAPEL's lexical network were given as input to the RP and AC methods. While the RP method exploited instances of the same type of the relation as the triples to ontologise, the AC method used the network established by relations held between at least one noun and another word, (e.g. hypernymy, part-of, producer-of, purpose-of).

As for other parameters, RP was first ran with a threshold $\theta = 0.5$, which guarantees higher precision. However, this value lead to low coverage rates, so we also ran RP with $\theta = 0.2$.

4.3 Coverage results

In tables 1, 3, 2, and 4 we present the coverage of the methods when ontologising the sample tb-triples in TePOT, computed with the following expression:

$$Coverage = \frac{|TriplesInSample|}{|OntologisedTriples|}$$

[4] Available through http://www.linguateca.pt/PAPEL/

The former tables show as well the number of ontologised triples according to their arguments: none of the arguments were in TePOT, so they were both attached to newly created single-word synsets (A); one argument attached to a new synset and the other to an existing synset (B); and both arguments attached to TePOT synsets (C).

When considering all the ontologised triples, the coverage of Random, AF and AC is close to 100% because, as referred in section 3, when TePOT does not contain the term in the argument of a triple, a new synset with that term is created. Still, none of the methods accept triples with both arguments attached to the same synset. When this happens, the tb-triple is not ontologised, which might lead to coverages slightly lower than 100% for Random and AF.

Furthermore, none of the methods ontologise triples when there is a tie in the selection of one argument. For RP, the triples are also not ontologised when one of the arguments has a related proportion lower than θ, so its has lower coverages, which, as expected, increase as θ decreases. It can also be observed that, despite being lower than for the other methods, RP's coverage for hypernymy is higher than for the other relations. This happens because the provided lexical network contains about 70,000 hypernym-of triples, but only about 2,300, 2,800 and 1,800 part-of, member-of and purpose-of triples respectively. Therefore, the odds of finding a proportion of related terms in a synset higher than θ are lower for the latter relations.

Coverage is presented both for all the tb-triples (Total), as well as considering only those whose arguments were attached to synsets already in TePOT (TS) and not new single-word synsets. The main reason for the differences of the coverage values is the coverage of TePOT itself. For instance, TeP, which is much larger than OT, is a broad-coverage thesaurus made for Brazilian Portuguese, while PAPEL contains triples extracted from an European Portuguese dictionary. Moreover, some authors [23] [26] have noticed that, even though the three aforementioned resources are broad-coverage Portuguese resources, they are more complementary than overlapping.

Method	**Ontolged. triples**	**Args.** A B C	**Cover. (%)** Total	TS
Rand.	499/500	141 13 345	**99.80**	**69.00**
AF	496/500	141 13 344	99.60	68.80
RP 0.5	315/500	107 13 195	63.00	39.00
RP 0.2	382/500	124 13 245	76.40	49.00
AC	458/500	139 13 306	91.60	61.20

Table 1. Coverage for hypernym-of triples.

Method	**Ontolged. triples**	**Args.** A B C	**Cover. (%)** Total	TS
Rand.	199/199	77 14 108	**100.00**	**54.27**
AF	199/199	77 14 108	**100.00**	**54.27**
RP 0.5	87/199	40 14 33	17.40	6.60
RP 0.2	129/199	53 14 62	64.82	31.16
AC	188/199	75 14 99	94.47	49.75

Table 2. Coverage for part-of triples.

Method	**Ontolged. triples**	**Args.** A B C	**Cover. (%)** Total	TS
Rand.	436/436	103 7 326	**100.00**	**74.77**
AF	427/436	103 7 317	97.94	72.71
RP 0.5	156/436	51 7 98	35.78	22.48
RP 0.2	266/436	21 7 238	61.01	54.59
AC	400/436	100 7 293	91.74	67.20

Table 3. Coverage for member-of triples.

Method	**Ontolged. triples**	**Args.** A B C	**Cover. (%)** Total	TS
Rand.	125/125	34 4 87	**100.00**	**69.60**
AF	125/125	34 4 87	**100.00**	**69.60**
RP 0.5	44/125	14 4 26	35.20	20.80
RP 0.2	73/125	22 4 47	58.40	37.60
AC	121/125	34 4 83	96.80	66.40

Table 4. Coverage for purpose-of triples.

5 Evaluation

Both manual evaluation of the resulting sb-triples, or the comparison to a gold resource with of all the correct sb-triples are possible approaches for assessing our results. Regarding that we do not have a gold resource available, we would have to create one manually. Therefore, besides relying too much on human judgements, both evaluation approaches would involve time-consuming tasks.

So, our first evaluation approach is based on support found automatically in a corpus and presents an alternative to the former approaches. It is easily repeatable and does not rely so much on human labour. Still, in order to complement the automatic approach, we ended up creating a small gold resource with all the acceptable sb-triples, given a set of tb-triples. Our choice was based on the fact that, after the creation of the resource, this evaluation approach could be repeated as many times as needed. Besides, this resource can be augmented and used in future evaluations.

In this section, we describe two different perspectives on the performance of the ontologisation methods: (i) automatic validation based on support found in a corpus (section 5.1); (ii) and gold resource evaluation (section 5.2). Their results are next reported and discussed.

5.1 Corpus-based validation

The resulting sb-triples were validated automatically, taking advantage of the text of CETEMPúblico [24], a syntactically annotated corpus with approximately 180 million words. We started by removing unfrequent words from our synsets, more precisely words with corpus frequency lower than 25. As dictionaries and thesauri have several words not frequently used in newspaper text, using the latter would have been an additional source of noise in the automatic comparison of performance. Then, we searched automatically in the corpus for support for each sb-triple. Finally, precision, recall and F_1 were estimated, providing the comparison of methods.

Validation metrics This approach is based on a set of lexico-syntactic patterns that typically denote semantic relations. The discriminating patterns were inspired by those used to find support for the triples of a former version of PAPEL[5][6].

For each synset, S_a and S_b, in a sb-triple, $S_a\ R\ S_b$, we searched for instances of each pair of terms $a_i \in S_a$ and $b_j \in S_b$ connected by a pattern denoting the relation R. If support was found at least once, $found(a_i, b_j, R) = 1$, otherwise

[5] The lexico-syntactic patterns are available through the website of a recent system that aims at the validation of Portuguese tb-triples, VARRA, http://www.linguateca.pt/VARRA/

[6] Despite of having two kinds of meronymy triples in PAPEL 2.0, part-of and member-of, the same discriminating patterns were used to validate both of them because these relations can be expressed by very similar ways (on this problem, see [12])

$found(a_i, b_j, R) = 0$. Finally, the approximate precision of each synset is given by the average validation of all the pairs, using the following expression:

$$Precision = \frac{\sum_{i=1}^{|S_a|} \sum_{j=1}^{|S_b|} found(a_i, b_j, R)}{|S_a| \times |S_b|}$$

Based on the precision, we can roughly estimate the recall and compute the F_1 measure:

$$Recall = \frac{Precision \times |OntologisedTriples|}{|TriplesInSample|} \qquad F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Tb-triples	**Sb-triples**
aparelho (device) **Hypernym-of** *televisor* (television)	*apresto,utensílio,petrechos,instrumento,apetrechos,aparelho* (tool,instrument,paraphernalia,device) **Hypernym-of** *tv,televisão,televisor,têvê* (tv,television)
extensão (extension) **Hypernym-of** *território* (territory)	*superfície,dimensão,extensão;espaço,área* (surface,dimension,extension,space,area) **Hypernym-of** *território,área* (territory,area)
ângulo (angle) **Part-of** *triângulo* (triangle)	*ângulo,face,lado* (angle,side,edge) **Part-of** *triângulo,trilateral* (triangle,trilateral)
técnica (technique) **Member-of** *marketing* (marketing)	*arte,técnica* (art,technique) **Member-of** *marketing* (marketing)
edição (edition) **Purpose-of** *programa* (program)	*edição,lançamento* (edition,launch) **Purpose-of** *programa,aplicativo* (program,application)

Table 5. Examples of top-scored triples, ontologised with AC

Examples of ontologised triples In table 5, we present some examples of high scoring triples of each type, ontologised using the AC method, which, as we will show in section 5.1, got the best validation results. Table 6 has more examples of triples ontologised with AC. Even though acceptable, the former did not get the top score. To complement the examples, we present the following sentences, which support some of the sb-triples in the aforementioned tables:

– *...as máquinas fotográficas digitais, os **televisores**, os videogravadores e outros **aparelhos** da electrónica de consumo...*
(...digital cameras, televisions, videorecorders and other electronic devices...)

- *...também não consigo reparar **televisões** e outros **aparelhos**.*
  (... nor can I repair televisions and other devices.)
- *A legislação existente define **superfície** agrícola como toda a **área** que...*
  (The current legislation defines agricultural land as the entire area that...)
- *O **ângulo** externo do **triângulo** (n = 3) se obtém dividindo 360 por 3.*
  (The external angle of the triangle (n = 3) is obtained after dividing 360 by 3.)
- *...prosseguindo assim o seu **programa** espacial, que tem como objectivo o **lançamento** de naves...*
  (... going on with their spacial program, which has the objective of launching spaceships...)
- *Peyroteo construiu um **palco** feito de muitos **palcos**.*
  (Peyroteo built a stage built of several stages.)
- *E não estão suficientemente precavidos contras as técnicas de marketing.*
  (And they are not wary of the marketing techniques.)
- *A guerra entre **seres** da mesma **raça** tornou-se no maior cataclismo jamais conhecido da história do Universo.*
  (War between beings of the same race became the greatest catastrophe in the history of the Universe.)
- *...é um disparate queimar **combustível** para produzir **calor**.*
  (...it is nonsense to burn fuel to produce heat.)

Tb-triples	Sb-triples	Score
terra (land) **Hypernym-of** *país* (country)	*plaga,lugar,região,terreno,terra* (place,region,terrain,land) **Hypernym-of** *território,pátria,país,região,nação* (territory,country,homeland,region,nation)	0.6
arte (art) **Hypernym-of** *escultura* (sculpture)	*arte,obra* (art,work) **Hypernym-of** *escultura,vulto,imagem,estátua* (sculpture,figure,image,statue)	0.75
cena (scene) **Part-of** *teatro* (theatre)	*cenário,painel,palco,panorama,cena* (scenary,panel,stage,picture,scene) **Part-of** *escultura,vulto,imagem,estátua* (theatre,stage,amphitheatre)	0.6
província (province) **Part-of** *país* (country)	*província,distrito,circunscrição,região* (province,district,division,region) **Part-of** *território,pátria,região,nação* (territory,homeland,region,nation)	0.55
pessoa (person) **Member-of** *raça* (race)	*ser,pessoa,criatura,indivíduo* (being,person,creature,individual) **Member-of** *raça,gente* (race,people)	0.5
calor (heat) **Purpose-of** *combustível* (fuel)	*calor,aquecimento,animação,entusiasmo* (heat,warmth,liveliness,enthusiasm) **Purpose-of** *combustível* (fuel)	0.5

Table 6. More examples of triples ontologised with AC

 ISBN: 978-989-95618-4-7

Validation results Tables 7, 8, 9 and 10 show the results of the automatic validation of the sb-triples obtained after ontologising hypernymy, part-of, member-of and purpose-of tb-triples respectively. The values of the validation metrics, introduced in section 5.1, are shown for all the ontologised triples (Total), including the ones attached to newly created single-word synsets, and also considering only triples with both arguments attached to TePOT synsets (TS). To minimise the impact of (hardly supported) triples connecting very unfrequent terms, before this validation, we removed from TePOT all the terms with less than 25 occurrences in the corpus.

Method	**Prec. (%)**		**Rec. (%)**		F_1 **(%)**	
	Total	TS	Total	TS	Total	TS
Rand.	33.46	27.37	33.40	18.88	33.43	22.35
AF	35.76	29.88	**35.62**	**20.56**	35.69	**24.36**
RP 0.5	**43.22**	**33.10**	27.23	12.91	33.41	18.58
RP 0.2	41.50	31.90	31.71	15.63	**35.95**	20.98
AC	36.45	29.16	33.39	17.85	34.85	22.14

Table 7. Validation of hypernym-of triples.

Method	**Prec. (%)**		**Rec. (%)**		F_1 **(%)**	
	Total	TS	Total	TS	Total	TS
Rand.	36.54	24.05	36.54	13.05	36.54	16.92
AF	37.30	25.05	**37.30**	13.60	37.30	17.63
RP 0.5	**52.99**	**28.55**	9.22	1.88	15.71	3.54
RP 0.2	47.71	28.43	30.93	8.86	37.53	13.51
AC	42.03	29.36	39.71	**14.60**	**40.83**	**19.51**

Table 8. Validation of part-of triples.

Method	**Prec. (%)**		**Rec. (%)**		F_1 **(%)**	
	Total	TS	Total	TS	Total	TS
Rand.	32.17	27.35	32.17	20.45	32.17	23.41
AF	36.50	29.76	**35.75**	**21.64**	**36.12**	**25.06**
RP 0.5	**43.64**	**33.04**	15.61	7.43	23.00	12.13
RP 0.2	41.98	32.66	25.61	17.83	31.82	23.06
AC	37.05	30.69	33.99	20.62	35.45	24.67

Table 9. Validation of member-of triples.

Method	**Prec. (%)**		**Rec. (%)**		F_1 **(%)**	
	Total	TS	Total	TS	Total	TS
Rand.	29.40	18.30	29.40	**12.74**	29.40	**15.02**
AF	29.14	17.54	29.14	12.21	29.14	14.40
RP 0.5	**49.73**	**23.63**	17.51	4.91	25.90	8.14
RP 0.2	40.69	21.62	23.77	8.13	30.01	11.81
AC	31.98	18.21	**30.96**	12.09	**31.36**	14.53

Table 10. Validation of purpose-of triples.

The results show that RP 0.5, followed by RP 0.2, is the most precise method for all relation types. So, for situations where precision is very important, RP with a high θ should be used. However, RP has low recall and thus low F_1.

The best F_1 is different, depending on the type of relation. AC is the best method for part-of and purpose-of, if all the sb-triples are considered. Even though it is not the best for the rest of the relations, the F_1's of the AC method are, most of the times, just outperformed by AF. As a frequency-based baseline, AF's performance revealed to be hard to beat. It benefited from the frequency of the terms in the synsets it connected, especially for hypernymy and member-of, where it obtained the highest F_1 for TS sb-triples.

For hypernymy, all measures are close to each other. Still, the best F_1 is obtained by RP 0.2, if all sb-triples are considered, and for AF, considering only TS sb-triples. On the other hand, member-of F_1 values are more oscilating, but higher for AF and AC. For purpose-of, if only TS sb-triples are considered, none of the methods outperform the Random baseline.

These results should not be viewed as completely conclusive, not only because RP and AC rely heavily on the information provided by the lexical network, but also because they are based on corpus support. While related terms can occur in

corpora connected by a huge amount of variations, the discriminating patterns used are only the most frequent and do not encompass all possibilities. Furthermore, some of them are ambiguous (e.g. *de*|*do*|*da*, 'of' in English, for meronymy), and may sometimes denote different semantic relations. Together with the high granularity of some senses and low ambiguity of some words in TePOT, the aforementioned situations increase the noise in this evaluation, which favours the precision of the Random and AF baselines.

5.2 Gold resource evaluation

In order to have a different perspective on evaluation, we ended up creating a gold resource. Although not large enough to make strong conclusions, this evaluation approach should be viewed as a complement to the automatic validation.

Gold resource creation The samples used in the experimentation (section 4) were the starting point of a gold resource. From them, we selected smaller subsets, considering only tb-triples whose attachment to TePOT synsets was acceptable and raised no doubts, more precisely 135 hypernym-of triples, 72 part-of triples, 105 member-of triples and 60 purpose-of triples. Then, we defined, manually, all the possible sb-triples for each tb-triple, regarding that, for some tb-triples, more than one attachment for one or both arguments made sense.

Table 11 shows the average number of possible sb-triples, in the gold resource, for each tb-triple in the samples ($\overline{Possible}$), the average number of correct sb-triples for each tb-triple ($\overline{Correct}$), and the proportion of correct sb-triples for each tb-triple ($\overline{Correct\%}$), according to the type of relation.

Relation	$\overline{Possible}$	$\overline{Correct}$	$\overline{Correct\%}$
Hypernym-of	14.09	3.92	42.05%
Part-of	9.54	2.72	45.49%
Member-of	12.61	4.58	52.40%
Purpose-of	12.53	4.95	52.47%

Table 11. Matching possibilities in the gold resource.

Method	**Hypernym-of** (135 tb-triples)			**Part-of** (72 tb-triples)			**Member-of** (105 tb-triples)			**Purpose-of** (60 tb-triples)		
	P (%)	R (%)	F_1(%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
Random	42.54	10.76	17.19	49.30	17.90	26.22	56.19	12.27	20.14	52.54	10.44	17.42
AF	43.18	10.78	17.25	45.83	16.84	24.63	52.48	11.02	18.21	46.55	9.09	15.21
RP 0.5	60.81	8.50	14.93	**66.67**	8.16	14.54	53.13	3.53	6.63	**68.75**	3.70	7.03
RP 0.2	57.45	10.21	17.34	58.70	13.78	22.31	56.25	7.48	13.21	59.38	6.40	11.55
AC	**61.16**	**13.40**	**22.77**	62.50	**20.41**	**30.77**	**70.00**	**13.10**	**22.07**	59.26	**10.77**	**18.23**

Table 12. Gold resource evaluation results.

Evaluation results The results of this evaluation approach, presented in table 12, reinforce the position of AC as the most accurate method for all relations, including not only part-of and member-of, as in the automatic validation, but also hypernymy and purpose-of. In opposition to the corpus-based evaluation, where it is favoured by the frequencies of the synsets, here, besides AC for all relations, AF it is outperformed by the Random baseline for part-of, member-of and purpose-of relations.

Moreover, high precision and low recall of RP 0.5 are still shown. For instance, for part-of and purpose-of, the precision of RP 0.5 outperforms AC's. For AC, RP 0.2 and AF, F_1's are lower for purpose-of triples. Besides the small sample of 60 tb-triples, this can be explained by the definition of the purpose-of relation in PAPEL. It can relate an object or a process to their purpose, or the goal they are used to achieve, which can be, for instance, a new state or an object.

As expected, the precision of the Random baseline is very close to the proportion of correct sb-triples in table 11, for hypernymy-of and purpose-of. For the other relations, it is some points higher without, however, any impact on the comparison. Finally, it is worth noticing that, since the ontologising methods only attach a term in a tb-triple to one synset, the gold resource might contain several possible sb-triples that are not matched, leading to recall rates of 20% or less.

5.3 Parallelism with related work

Besides RP following the same assumption as the anchor approach [19], we did not consider other methods referred in section 2. While these methods rely on other relations or glosses in WordNet, the freely available thesaurus for Portuguese are just synset bases which do not provide relations nor glosses. Furthermore, some methods [25] take advantage of the extraction context, which we assumedly do not want to use.

Also, our experimentation cannot be compared to [19]'s, not only because our algorithms did not use existing hypernymy sb-triples but, especially because our work was done for Portuguese, using a thesaurus which is not as large and broad as Princeton WordNet is for English.

We should finally add that AF is different from hard to beat frequency-based baselines [6] using Princeton WordNet, as our synset frequencies were roughly estimated, based on the frequency of their terms in Google. So, while the frequency of the terms in the synsets connected with AF favoured its overall performance in the corpus-based validation, its performance dropped in the gold standard evaluation. In order to have a ranking similar to WordNet's, we would first need to annotate the senses of the words in a corpus, according to a Portuguese synset inventory, in a similar fashion to SemCor [14], for English.

6 Concluding remarks

We have presented several methods for ontologising tb-triples into a synset base using only extracted triple sets and not the extraction context. It is our intention

to use at least one of these methods in a broader system that aims the automatic construction of a lexical ontology for Portuguese [8], so the comparison of their accuracy is important for us. We believe that the results obtained are interesting, but there is still a long way to go, since this is a very challenging task.

Considering evaluation, the gold resource should be enlarged in the future, the manual attachments should be made by more than one person and their agreement should be measured. Furthermore, methods for improving TePOT will be devised, including its augmentation with information extracted from electronic dictionaries. Also, strategies for increasing the recall include having the possibility of establishing more than one sb-triple per tb-triple.

Acknowledgements

Hugo Gonçalo Oliveira is supported by the FCT scholarship grant SFRH/BD/44955/2008, co-funded by FSE.

References

1. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proc. 5th ACM Intl. Conf. on Digital Libraries. pp. 85–94 (2000)
2. Agirre, E., Lacalle, O.L.D., Soroa, A.: Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In: Proc. 21st Intl. Joint Conf. on Artifical Intelligence (IJCAI). pp. 1501–1506. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2009)
3. Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proc. 37th Annual Meeting of the ACL. pp. 120–126. ACL Press, Morristown, NJ, USA (1999)
4. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM 51(12), 68–74 (2008)
5. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MIT Press (1998)
6. Gale, W.A., Church, K.W., Yarowsky, D.: One sense per discourse. In: Proc. HLT'91 workshop on Speech and Natural Language. pp. 233–237. ACL Press, Morristown, NJ, USA (1992)
7. Gale, W.A., Church, K.W., Yarowsky, D.: Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In: Proc. 30th Annual Meeting of the ACL. pp. 249–257. ACL Press, Morristown, NJ, USA (1992)
8. Gonçalo Oliveira, H., Gomes, P.: Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In: Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010). IOS Press (2010)
9. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Relations extracted from a portuguese dictionary: results and first evaluation. In: Local Proc. 14th Portuguese Conf. on Artificial Intelligence (EPIA) (2009)
10. Harabagiu, S.M., Moldovan, D.I.: Enriching the wordnet taxonomy with contextual knowledge acquired from text. In: Natural language processing and knowledge representation: language for knowledge and knowledge for language, pp. 301–333. MIT Press, Cambridge, MA, USA (2000)

11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. 14th Conf. on Computational Linguistics. pp. 539–545. ACL Press, Morristown, NJ, USA (1992)
12. Ittoo, A., Bouma, G.: On learning subtypes of the part-whole relation: Do not mix your seeds. In: Proc. 48th Annual Meeting of the ACL. pp. 1328–1336. ACL Press, Uppsala, Sweden (2010)
13. Maziero, E.G., Pardo, T.A.S., Felippo, A.D., Dias-da-Silva, B.C.: A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do brasil. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL). pp. 390–392 (2008)
14. Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: Proc. ARPA Human Language Technology Workshop. Plainsboro, NJ, USA (1994)
15. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys 41(2), 1–69 (2009)
16. Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Extending and enriching Wordnet with OntoLearn. In: Proc. 2nd Global WordNet Conf. (GWC). pp. 279–284. Masaryk University, Brno, Czech Republic (2004)
17. Pantel, P.: Inducing ontological co-occurrence vectors. In: Proc. 43rd Annual Meeting of the ACL. pp. 125–132. ACL Press (2005)
18. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proc. 21st Intl. Conf. on Computational Linguistics and 44th Annual Meeting of the ACL. pp. 113–120. ACL Press, Sydney, Australia (2006)
19. Pennacchiotti, M., Pantel, P.: Ontologizing semantic relations. In: Proc. 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the ACL. pp. 793–800. ACL Press (2006)
20. Ponzetto, S.P., Navigli, R.: Knowledge-rich word sense disambiguation rivaling supervised systems. In: Proc. 48th Annual Meeting of the ACL. pp. 1522–1531. ACL Press, Uppsala, Sweden (2010)
21. Resnik, P.: Disambiguating noun groupings with respect to Wordnet senses. In: Proc. 3rd Workshop on Very Large Corpora. pp. 54–68 (1995)
22. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: Proc. Advances in Web Intelligence, 3rd Intl. Atlantic Web Intelligence Conf. (AWIC). pp. 380–386. Springer (2005)
23. Santos, D., Barreiro, A., Freitas, C., Oliveira, H.G., Medeiros, J.C., Costa, L., Gomes, P., Silva, R.: Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In: Brito, A.M., Silva, F., Veloso, J., Fiéis, A. (eds.) Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística, pp. 681–700. APL (2010)
24. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proc. 39th Annual Meeting of the ACL. pp. 442–449. ACL Press (9-11 July 2001)
25. Soderland, S., Mandhani, B.: Moving from textual relations to ontologized relations. In: Proc. AAAI Spring Symposium on Machine Reading (2007)
26. Teixeira, J., Sarmento, L., Oliveira, E.: Comparing verb synonym resources for portuguese. In: Proc. Computational Processing of the Portuguese Language, 9th Intl. Conf. (PROPOR 2010). LNAI, vol. 6001, pp. 100–109. Springer (2010)