 ISBN: 978-989-95618-4-7

Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary

Leticia Antón Pérez[1,2], Hugo Gonçalo Oliveira[1], and Paulo Gomes[1]

[1] CISUC, University of Coimbra, Portugal,
[2] Higher School of Computer Engineering, University of Vigo, Spain,
leticiaap86@gmail.com, {hroliv,pgomes}@dei.uc.pt

Abstract. Public domain collaborative resources like Wiktionary and Wikipedia have recently become attractive sources for information extraction. To use these resources in natural languague processing (NLP) tasks, efficient programmatic access to their contents is required. In this work, we have extracted semantic relations automatically from the Portuguese Wiktionary and compared our results with the relations in PAPEL, a public domain lexical network extracted from a proprietary dictionary. We have found about 44,000 relations that were not in PAPEL, which suggests that Wiktionary is a valuable alternative source for enriching existing lexical knowledge bases.

Keywords: Collaborative Knowledge Bases, Information Extraction, Lexical-Semantic Knowledge.

1 Introduction

Today, it is increasingly easy to find large amounts of information on almost any topic. Amongst the large number of facilities that the World Wide Web provides, collaborative resources allow everyone to share information, which is causing a significant growth of semi-structured information. The aforementioned resources, such as Wikipedia and Wiktionary, are becoming popular alternatives for exploitation in the automatic creation of knowledge bases, and are thus used in various NLP tasks [11].

Even though the Wiktionary has not attracted as many researchers as Wikipedia, electronic dictionaries have been used since the 1980s as a primary target on the acquisition of lexical knowledge and automatic construction of lexical knowledge bases [3] [15] [13] [7]. Therefore, it seems natural to use the Wiktionary as an additional source of lexical knowledge which, in opposition to static dictionaries, is freely available for researchers and will keep growing over the next years.

In this paper, we exploit the Portuguese Wiktionary to extract semantic relations and show how it can contribute to the amplification of a public domain lexical network for Portuguese, namely PAPEL [7], extracted automatically from a proprietary dictionary.

We start by introducing work related to the automatic extraction of lexical-semantic knowledge and on using collaborative resources for information extraction. In section 3, we present how to access the contents of the Portuguese Wiktionary and how its contents are transformed into a more friendly format. Section 4 describes the process for extracting lexical semantic relations and reports on the results obtained after the whole process. Section 5 shows a validation approach of some of the results, based on support found in a corpus. Finally, section 6 provides a comparison with PAPEL considering the words and semantic relations covered by both resources.

2 Related Work

Dictionaries are an obvious target for the automatic extraction of lexical-semantic knowledge, as they are structured on words and meanings. Early works [2] [1] studied the regularity of dictionary entries and analysed how these resources could be exploited in the extraction of semantic relations, useful for the creation of large semantic networks. Following the aforementioned works, the first automatic procedures for creating taxonomies from dictionaries were developed [3].

Although dictionaries were used with several goals, such as word sense disambiguation or parsing, to our knowledge MindNet [15] was the first independent lexical knowledge base automatically created from dictionaries. Besides the construction of a broad-coverage knowledge base, the goal of MindNet is to provide a set of tools for acquiring, structuring, assessing and exploiting semantic information from natural language text, particularly from dictionaries.

In contrast to the research described above, several problems about semantic networks extracted from dictionaries have been pointed out [10]. One important conclusion is that a broad-coverage knowledge base requires the combination of multiples dictionaries. This idea arises because the information contained in these resources differs considerably in amount and kind.

After the establishment of Princeton WordNet [4] as a widely used public domain lexical resource for English, less attention has been given to information extraction from dictionaries. Nevertheless, there are recent works on the acquisition of ontologies from dictionaries [13] and on the extraction of relations, including, for Portuguese, PAPEL [7], a lexical network with relations automatically extracted from the definitions of a proprietary Portuguese dictionary.

Besides dictionaries, there are other popular sources of knowledge, such as corpora [8] or collaborative encyclopedias, used either to create larger resources or to enrich existing knowledge bases, such as WordNet. Today, the collaborative encyclopedia Wikipedia[3] is one of the most attractive sources for information extraction and has been used in numerous NLP and knowledge discovery tasks [11], including the acquisition of lexical-semantic knowledge [9] [22] [5].

[3] http://www.wikipedia.org/

Besides Wikipedia, the Wikimedia Foundation[4] runs Wiktionary[5], a multilingual project, which provides electronic dictionaries of free content, with definitions, examples, and information on part-of-speech (POS), translations, pronunciation and etymology, as well as information on semantic relations, such as synonyms, antonyms or hypernyms. However, as Wikipedia and Wiktionary are built manually by non-professional volunteers on the Web, some of the aforementioned pieces of information are not always complete and are sometimes inconsistent. On the other hand, Wiktionary is free and always growing. It contains approximately 5 million entries in 170 language editions[6]. In opposition to common dictionaries, Wiktionary has entries in other languages distinct of their own editing language. For example, the Portuguese Wiktionary has entries in English, Spanish or Italian.

So far, Wiktionaries have been exploited for computing semantic relatedness [23] [21] , useful for solving word choice problems, and synonymy networks have been extracted from Wiktionaries and compared to synonymy networks from other resources [12]. Moreover, Wiktionary has been used for the enrichment of lexical resources [17] as well as in the automatic creation of new lexical resources [20] [16].

3 Data Preprocessing

This section describes how to use the Portuguese Wiktionary XML dump to convert the Wiktionary definitions to a more friendly format, suitable for being parsed. Then, the vocabulary used in the definitions is analysed in order to see how it can be exploited. Some of the most frequent n-grams in the definitions are presented together with relations they may denote.

3.1 Processing the XML Dump and Exporting the Definitions

There are tools developed for providing the programmatic access to information in collaborative knowledge bases. For instance, there are two application programming interfaces (APIs) for the English and German Wikipedia and Wiktionary [22], especially designed for mining the lexical-semantic information dispersed in these knowledge bases, and to provide efficient and structured access to the available knowledge. However, the former Wiktionary API does not work for the Portuguese Wiktionary and, since the source code is not provided, we could not use it as a starting point. Therefore, we decided to develop an API for the Portuguese Wiktionary from scratch.

We have used our API (which will soon be in the public domain) to access programmatically some of the information contained in the Portuguese Wiktionary. The API works on the top of freely available Wiktionary XML dumps[7],

[4] http://wikimediafoundation.org/
[5] http://www.wiktionary.org/
[6] http://meta.wikimedia.org/wiki/Wiktionary, on May 2011
[7] http://dumps.wikimedia.org/

which contain the entries in the Portuguese Wiktionary and its contents. As in other projects of the Wikimedia Foundation, Wiktionary content is based on MediaWiki software[8] and wikitext.

After parsing the XML structure of the dump, the API we needed to analyse the structure of the entries in order to create a parser for the wikitext. Different language editions of Wiktionary use distinct delimiter elements to represent the information of each entry, so every Wiktionary parser needs to be adapted according to the language edition. Another important aspect, as noted earlier, is that there is no "formal syntax defined" for wikitext. Therefore, users can express the same idea in different ways. As no user is forced to follow the standard, the structure of the entries is fairly inconsistent, and, as a consequence, it is not possible to build a 100% reliable parser [12].

At the technical level, to build the parser we used several regular expressions to extract the contents of each entry as well as to eliminate mark-up which was not necessary. Once built, the parser is integrated in the API and is used to extract all entries of the Portuguese Wiktionary that refer to words in the Portuguese language. For this work we used the 24th April 2011 XML dump of the Portuguese Wiktionary. It contains over 170,000 entries, of which about 110,000 refer to Portuguese words.

From the latter entries, we exported all definitions we could to a more friendly data format where each line contains the head word, its part-of-speech and a definition. Entries with more than one definition gave rise to more than one line in our data format. Also, some of the entries of the Portuguese Wiktionary have synonymy lists, which gave rise to definitions. On the other hand, since semantic relations are established between open category words, more precisely nouns, verbs, adjectives and adverbs, we did not export definitions of words belonging to other parts-of-speech, including for instance, prepositions or inflected verbs[9]. Moreover, due to parsing issues related to inconsistencies in the wikitext, several entries do not origin any definition.

Figure 1 is an example of how information is structured in the XML dump for an entry of Portuguese Wiktionary and figure 2 shows the resulting definitions obtained in our data format. In the end, more than 66,000 definitions were collected, distributed according to table 1, where an example of a definition of each part-of-speech is given.

3.2 Analysing the Vocabulary of the Definitions

In order to identify frequent patterns used in the Portuguese Wiktionary, with special attention to those that denote semantic relations, we analysed the vocabulary used in the definitions. Frequent patterns are candidates for being exploited and thus included in grammars used for extracting semantic relations.

To this end, we compiled the n-grams (with $n = 2, 3, 4, 5$ and 6) in a table according to the part-of-speech of the definitions they occurred at, together with

[8] http://www.mediawiki.org/

[9] Some verbs have different entries for different verbal forms.

```
<page>
<title>computador</title>
...
<text xml:space="preserve">wikipedia
=Português=
==Adjetivo==
flex.pt|ms=computador|mp=computadores|fs=computadora|fp
=computadoras
oxítona|com|pu|ta|dor
# que [[computar|computa]]
==Substantivo==
flex.pt|ms=computador|mp=computadores
oxítona|com|pu|ta|dor
# o que [[fazer|faz]] [[cmputo]]s ([[cálculo]]s);
o que [[computar|computa]]
# [[máquina]] [[capaz]] de [[fazer]] [[cálculo]]s
===Sinônimos===
De '''1''': [[calculista]]
De '''2''': [[calculadora]]
===Tradução===
tradini|De 3 (aparelho eletrônico capaz de calcular)
trad|af|rekenaar
...
</page>
```

Fig. 1. Information in the XML dump for the entry *computador* (computer) in the Portuguese Wiktionary

computador	adj	que computa
computador	nome	o que faz cômputos (cálculos); o que computa
computador	nome	máquina capaz de fazer cálculos
computador	adj	calculista
computador	nome	calculadora

Fig. 2. Definitions obtained from the entry in figure 1

their frequency, which would be useful for taking conclusions on their utility for the extraction of semantic relations. Some of the more interesting and frequent n-grams and their rough translations are shown in table 2, along with their frequency, the part-of-speech of the definitions they occur at and the type of relation they usually denote.

After observing the most frequent n-grams, we concluded that most of them were already covered by the grammars used for the creation of PAPEL, freely available through http://www.linguateca.pt/PAPEL. Therefore, we decided to use the aforementioned grammars for extracting the same kinds of relations, and added just a few missing rules to increase recall.

4 Extraction

This section describes the procedure for the automatic extraction of semantic relations and presents the results obtained with this procedure.

Table 1. Number of definitions according to their part-of-speech

POS	Definitions	Examples
Nouns	41,836	*homem: um tipo de primata bípede e bímano da espécie Homo Sapiens* (man: a kind of primate bipedal and bimane of the Homo Sapiens species)
Verbs	8, 703	*avermelhar: fazer ou tornar vermelho* (redden: to make or to become red)
Adjectives	14,987	*bípede: que tem dois membros* (biped: with two members)
Adverbs	909	*moralmente: de maneira moral* (morally: in a moral way)

4.1 Procedure

After collecting the definitions, semantic relations are extracted through a manual stage and two automatic stages. This procedure was inspired by the construction of PAPEL, reported in [7].

Manual creation of the extraction grammars: during the analysis of the vocabulary, several productive patterns arise and are thus compiled in grammars, especially made for the extraction of relations between words in the definitions and head words. However, as referred in section 3.2, we reused the grammars originally made for the creation of PAPEL.

Extraction of semantic relations: in this stage, the grammars are used together with a parser, PEN[10] that processes the definitions in the data format introduced in section 3.1. In the end, if definitions match the patterns, instances of semantic relations are extracted. These relations are denoted by relational triples $t = (a, R, b)$ where, for instance, a is a word in the definition, b is the head word, and R is the name of the relation.

Adjustment of relations and lemmatization: after the extraction, some relations have invalid arguments, including punctuation marks or prepositions. Besides discarding the former relations, definitions are part-of-speech tagged with the POS-tagger provided by OpenNLP[11], using the models for Portuguese[12]. Then, based on the annotation of the definition and on the parts-of-speech provided by Wiktionary, some relation types are adjusted. If possible, the relation types are changed to a type of the same group, so that the part-of-speech of its arguments are matched. Otherwise, the triple is discarded. Finally, if the arguments of the triples are inflected, we apply some lemmatization rules to the arguments of the triples.

[10] PEN is the parser used in the construction of PAPEL, and can be downloaded from http://code.google.com/p/pen/
[11] http://incubator.apache.org/opennlp/
[12] http://opennlp.sourceforge.net/models-1.5/

Table 2. More interesting and frequent n-grams

N-gram	Frequency	Part-of-speech	Semantic relation
o mesmo que (the same as)	756	Noun	Synonymy
ato ou efeito de (act or effect of)	435	Noun	Causation
conjunto de (set of)	248	Noun	Member-of
pessoa que (person who)	237	Noun	Hypernymy
espécie de (species of)	175	Noun	Hypernymy
o mesmo que (the same as)	72	Verb	Synonymy
relativo à/ao (relative to)	619	Adjective	Property
que se (that)	399	Adjective	Property
que tem (that has)	382	Adjective	Part-of/Property
diz-se de (it is said about)	245	Adjective	Property
habitante ou natural de (inhabitant or natural of)	143	Adjective	Place/Origin
de modo (in a way)	82	Adverb	Manner
de maneira (in a manner)	26	Adverb	Manner

4.2 Results

Once the extraction process is applied to the definitions extracted from the Portuguese Wiktionary, 55,705 relational triples are obtained, distributed according to table 3. This table presents the groups of semantic relations, the name of the relation type considering the part-of-speech of the arguments, the number of relations extracted of each type and an example of each relation type extracted. As expected, synonymy and hypernymy are the relations with most triples. Regarding unique words in the arguments of relational triples, 22,369 nouns, 6,326 verbs, 6,854 adjectives and 465 adverbs were extracted

5 Automatic Validation

Since manual evaluation of this kind of knowledge is a time-consuming and typically subjective task, we followed an automatic validation procedure based on support found for the relations in a corpus. This procedure, also inspired by the

Table 3. Extracted semantic relations

Group	Name	Args.	Qnt.	Examples
Synonymy	SINONIMO_N_DE	n,n	13,647	*léxico,dicionário* (lexicon,dictionary)
	SINONIMO_V_DE	v,v	4,136	*esticar,estender* (to_extend,to_stretch)
	SINONIMO_ADJ_DE	adj,adj	6,660	*pronto,súbito* (prompt,sudden)
	SINONIMO_ADV_DE	adv,adv	292	*imediatamente,já* (imediatelly,now)
Hypernymy	HIPERONIMO_DE	n,n	18,638	*pessoa,guerreiro* (person,warrior)
Part-of	PARTE_DE	n,n	723	*núcleo,átomo* (core,atom)
	PARTE_DE_ALGO_COM_PROP	n,adj	559	*vício,vicioso* (addiction,addictive)
	PROPRIEDADE_DE_ALGO_PARTE_DE	adj,n	26	*sujeito,oração* (subject,sentence)
Member-of	MEMBRO_DE	n,n	1,166	*aluno,escola* (student,school)
	PROPRIEDADE_DE_ALGO_MEMBRO_DE	adj,n	121	*rural,campo* (rural,country)
	MEMBRO_DE_ALGO_COM_PROP	n,adj	23	*coisa,coletivo* (thing,collective)
Causation-of	ACCAO_QUE_CAUSA	v,n	1,149	*mover,movimento* (to_move,movement)
	CAUSADOR_DE	n,n	307	*vírus,doença* (virus,disease)
	PROPRIEDADE_DE_ALGO_QUE_CAUSA	adj,n	163	*horrível,horror* (horrible,horror)
	CAUSADOR_DA_ACCAO	n,v	5	*fogo,fundir* (fire,to_melt)
Producer-of	PRODUTOR_DE	n,n	316	*oliveira,azeitona* (olive_tree,olive)
	PROPRIEDADE_DE_ALGO_PRODUTOR_DE	adj,n	37	*explosivo,explosão* (explosive,explosion)
	PRODUTOR_DE_ALGO_COM_PROPRIEDADE	n,adj	8	*fermentação,fermentado* (fermentation,fermented)
Purpose-of	ACCAO_FINALIDADE_DE	v,n	1,485	*calcular,cálculo* (to_calculate,calculation)
	FINALIDADE_DE	n,n	1,355	*sustentao,mastro* (support,mast)
	ACCAO_FINALIDADE_DE_ALGO_COM_PROP	v,adj	27	*comprimir,compressivo* (to_compress,compressive)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	10	*habitação,habitável* (habitation,inhabitable)
Location	LOCAL_ORIGEM_DE	n,n	747	*Índia,hindu* (India,hindu)
Manner	MANEIRA_POR_MEIO_DE	adv,n	157	*estranhamente,estranho* (oddly,odd)
	MANEIRA_SEM	adv,n	24	*prontamente,demora* (promptly,delay)
	MANEIRA_SEM_ACCAO	adv,v	5	*seguido,parar* (straight,to_stop)
Property-of	PROPRIEDADE_DO_QUE	adj,v	3,256	*repousado,repousar* (restful,to_rest)
	PROPRIEDADE_DE_ALGO_REFERENTE_A	adj,n	1,635	*daltónico,daltonismo* (daltonic,daltonism)

one used to validate the relations of PAPEL, uses a set of discriminating patterns indicative of semantic relations and searches for instances of the extracted semantic relations in text.

We recall that the results obtained should not be confused with the precision of the extracted relations, but can be used as a quality indicator. First, because a corpus is a limited source of knowledge. Then, there are many ways to express a semantic relation in text and it is thus impossible to encode all the possible patterns. Last but not least, many relations are very specific and tend to occur only in dictionaries. For instance, it is not common to find relations as ACCAO_QUE_CAUSA, which implies the nominalisation of a verb (eg. in *aumentar* ACCAO_QUE_CAUSA *aumento*, in English 'augment' causes 'augmentation'), as their arguments are not likely to co-occur in corpora text. Also, synonymous words tend not to co-occur in corpora text, especially in the same sentence, as the writer tends to use always the same word for denoting the same concept. Therefore, only some semantic relations, all of them held between nouns, were validated. Still, if the same patterns and the same corpus are used in further evaluations, this metric can be used to compare resources based on relational triples.

In this validation, we have used the newspaper corpus CETEMPúblico [18], a syntactically annotated newspaper corpus, with approximately 180 million words. The list of discriminating patterns used as a starting point the patterns of a VARRA, a system that uses Portuguese corpora for finding examples of semantic relations in context. VARRA's web interface and the list of patterns are available through http://www.linguateca.pt/VARRA. In table 4 we present two perspectives on the results according to two different criteria:

- First, we selected only the triples whose arguments occurred more than 100 times in the corpus (Freq(args) $\geq$ 100).
- Then, we selected only the triples whose arguments co-occurred at least once in the same sentence (Cooc(args) $\geq$ 1).

Also in table 4, the columns 'Total' contain the number of triples matching the previous criteria and the proportion of the total number of triples of each relation they represent. The column 'Supported' presents the number of triples supported by the corpus and their proportion according to the triples matching the criteria. About one third of the hypernymy and part-of triples matching first criteria were supported. Slightly more member-of triples, and just about 15% of the causation and purpose-of triples were supported. The proportion of relations matching the second criteria is higher for all the relations but causation. The proportion of supported relations following the latter criteria is, as well, higher, this time for all relations. Even though not substantial, the difference between the proportion of supported relations in the both criterias is higher for hypernymy and part-of. This proportion is lower for purpose-of and causation-of relations because their discriminating patterns are more, have more variations, and are less common.

Table 5 contains examples of relational triples supported by the corpus and the sentences that support them, together with their rough translations. Patterns are in bold in the latter table.

Table 4. Results of the automatic validation

Relation	Freq(args) ≥ 100				Cooc(args) ≥ 1			
	Total		Supported		Total		Supported	
Hypernymy	6,556	35.2%	2,074	31.6%	6,584	36.9%	2,249	34.2%
Part-of	235	32.5%	91	38.7%	238	33.1%	99	41.6%
Member-of	323	27.7%	144	44.6%	329	28.6%	149	45.3%
Causation	99	32.2%	14	14.1%	82	26.8%	12	14.6%
Purpose-of	440	32.5%	66	15.0%	445	33.0%	70	15.7%

6 Comparison with PAPEL

As referred in section 2, PAPEL is a public domain lexical network for Portuguese with around 198,000 relational triples. Besides the possibility of reusing the grammars, one of the reasons that made us follow the PAPEL data format was the possibility of comparing it with our resource. Furthermore, it enables to merge these two resources in a broader resource, as suggested by several authors [10] to improve the quality of knowledge extracted from dictionaries.

Therefore, in table 6, we put the quantities of relational triples in the last version of PAPEL, 2.0, side-by-side with the quantity of the relational triples extracted from Wiktionary. In table 7, we do the same for unique words in relational triples. Both of these tables show as well the percentage of new relational triples/words Wiktionary has to offer to PAPEL, computed as the ratio between the relational triples/words from Wiktionary not in PAPEL (New) and relational triples/words already in PAPEL 2.0.

Even though PAPEL is much larger, there are more new relational triples extracted from Wiktionary not found in PAPEL 2.0 (44,201), than common triples (11,387), which reinforces the idea that these two resources can be merged in a lexical resource with higher coverage. Also, more than half of the words in relational triples from the Wiktionary are in PAPEL 2.0 (21,807), but there is still a great amount of new words (14,207). For verbs, according to the work of other authors [19], we were already expecting that the Wiktionary could provide an interesting amount of words that are not in PAPEL.

Finally, to compare the proportion of triples in PAPEL supported by CETEMPúblico with the number of triples extracted from Wiktionary in the same situation, we have validated the triples of PAPEL the same way as we had done for the triples extracted from Wiktionary (whose results are reported in section 5). Table 8 has the same contents as table 4, but for PAPEL. This table shows that there are more triples extracted from Wiktionary matching both criteria (arguments frequency higher than 100 and co-occurring arguments), than in PAPEL, which has lower proportions of supported triples as well. This can be explained because PAPEL is larger and was extracted from a commercial dictionary, created by lexicographers, who are experts on describing word senses. On the other hand, Wiktionary is smaller and is created by non-expert volunteers. Therefore, while PAPEL, besides including more common relational triples, has

Table 5. Sentences that support extracted relations

Relation	Support found
língua HIPERONIMO_DE *alemão* *language* hypernym_of *german*	*As iniciativas deste gabinete passam geralmente pela promoção de conferências, exposições, workshops e aulas de* ***línguas, como o inglês, alemão*** *ou japonês.* *The initiatives of this office are generally for the promotion of conferences, exhibitions, workshops and classes in languages like English, German or Japanese.*
ciência HIPERONIMO_DE *grafologia* *science* hypernym_of *graphology*	*Mas para Alberto Vaz da Silva,* ***a grafologia é uma ciência*** *que, além de definir o carácter e temperamento de um indivíduo, pode ajudá-lo a libertar-se de culpas e complexos ganhos na infância.* But for Alberto Vaz da Silva, graphology is a science that, besides defining the character and the temperament of an individual, can help him to free himself from guilt and complexes acquired during childhood.
rua PARTE_DE *quarteirão* *street* part_of *block*	*...* ***quarteirão formado pelas ruas*** *de Rodrigues de Freitas, dos Polacos e de Marciano Aziaga, ...* *... block formed by the streets Rodrigues de Freitas, dos Polacos and Marciano Aziaga, ...*
mercúrio PARTE_DE *amálgama* *mercury* part_of *amalgam*	*O mercúrio uma substância altamente tóxica e as* ***amálgamas dentrias são feitas de mercúrio.*** *Mercury is a highly toxic substance and dentary amalgams are made of mercury.*
pessoa MEMBRO_DE *comissão* *person* member_of *committee*	*A* ***comissão é constituída por pessoas*** *que ficaram marcadas pela presença de Dona Amélia: ...* *The committee consists of people who were marked by the presence of Dona Amélia: ...*
lobo MEMBRO_DE *alcateia* *wolf* member_of *pack*	*Mech e os seus colegas constataram que alguns dos cheiros contidos nas marcas de urina servem para os* ***lobos de uma alcateia*** *saberem por onde andou o lobo que deixou as marcas ...* *Mech and his colleagues found that some of the smells of urine contained in the marks are for a pack of wolves to know where the wolf that left the marks has been...*
transporte FINALIDADE_DE *embarcação* *transport* purpose_of *ship*	*... onde foi descoberto o resto do casco de uma* ***embarcação presumivelmente utilizada no transporte*** *de peças de cerâmica ...* *... where the rest of the hull of a ship, allegedly used to transport pieces of pottery, was discovered ...*
vírus CAUSADOR_DE *doença* *virus* causation_of *disease*	*A hepatite A transmite-se enquanto as pessoas não têm sintomas, é uma* ***doença benigna, provocada por um vírus*** *que causa fraqueza, incómodos, febre e vómitos ...* *Hepatitis A is spread when people have no symptoms, is a benign disease caused by a virus that causes weakness, discomfort, fever and vomiting ...*

other formal and less conventional triples, the relational triples from Wiktionary tend to be more conventional and thus more likely be supported in corpora.

Table 6. Numbers of relational triples in PAPEL 2.0 and triples extracted from the Portuguese Wiktionary

Semantic relation	Args	PAPEL	Wiktionary	Common	New
Hypernymy	n,n	62,591	17,837	3,442	14,395 (+23%)
Synonymy	n,n	37,452	13,556	2,949	10,607 (+28%)
	v,v	21,465	4,076	1,275	2,801 (+13%)
	adj,adj	19,073	6,629	1,740	4,889 (+26%)
	adv,adv	1,171	289	94	195 (+17%)
Part-of	n,n	2,805	718	47	671 (+24%)
	n,adj	3,721	558	146	412 (+11%)
	adj,n	17	26	0	26 (+153%)
Member-of	n,n	5,929	1,152	83	1,069 (+18%)
	n,adj	34	23	1	22 (+65%)
	adj,n	883	121	10	111 (+13%)
Causation	n,n	1,013	306	14	292 (+29%)
	adj,n	498	163	21	142 (+29%)
	v,n	6,399	1,139	645	494 (+8%)
	n,v	39	5	0	5 (+13%)
Producer-of	n,n	898	310	14	296 (+33%)
	n,adj	35	8	0	8 (+23%)
	adj,n	359	37	7	30 (+8%)
Purpose-of	n,n	2,886	1,349	46	1,303 (+45%)
	n,adj	63	9	2	7 (+11%)
	v,n	5,192	1,479	126	1,353 (+26%)
	v,adj	260	27	6	21 (+8%)
Place-of	n,n	849	722	1	721 (+85%)
Manner-of	adv,n	1,113	156	64	92 (+8%)
Manner-without	adv,n	117	24	7	17 (+15%)
	adv,v	11	5	1	4 (+36%)
Property-of	adj,v	17,543	3,239	404	2,835 (+16%)
	adj,n	6,518	1,625	242	1,383 (+21%)

7 Conclusions And Future Work

We have shown that the Portuguese Wiktionary is a valuable resource, suitable for exploitation in the automatic creation or enrichment of lexical knowledge bases. After converting the contents of Wiktionary to a more friendly format, several semantic relations were extracted, validated and compared to the relations of an existing lexical network. It was observed that a substantial amount of new words and relational triples can be obtained from the former resource.

Table 7. Comparison of words in PAPEL and extracted from Wiktionary

POS	PAPEL	Wiktionary	Common	New
Nouns	55,933	22,369	13,451	8,918 (40%)
Verbs	24,061	6,326	3,368	2,958 (47%)
Adjectives	21,001	6,854	4,677	2,177 (32%)
Adverbs	1,390	465	311	154 (33%)

Table 8. Automatic validation of PAPEL

Relation	Freq(args) $\geq$ 100				Cooc(args) $\geq$ 1			
	Total		Supported		Total		Supported	
Hypernymy	17,749	28.4%	5,196	29.3%	18,511	29.6%	5,749	31.1%
Part-of	625	22.3%	183	29.3%	666	23.7%	212	31.8%
Member-of	1,035	17.5%	445	43.0%	1150	19.4%	462	40.2%
Causation	227	22.4%	35	15.4%	218	21.5%	41	18.8%
Purpose-of	839	29.1%	126	15.0%	855	29.6%	134	15.7%

The work presented here is part of a larger project aiming at the creation of a broader lexical resource for Portuguese, Onto.PT [6], where other freely available textual resources, including not only dictionaries, but also Wikipedia, are being exploited. Concerning Wikipedia, in order to avoid the inclusion of too much specific encyclopedic knowledge, our idea is to select the most relevant Wikipedia entries for each word in the lexical network, and then extract semantic relations from the abstracts of these entries. One advantage of using collaborative resources is that these resources keep growing and so will the results obtained from their exploitation.

We would like to add a word on using machine learning techniques in our work. Even though these techniques are widely used in the acquisition of semantic relations from corpora, when it comes to dictionary definitions, we believe that they are not the best alternative. Preliminary results obtained after using a weakly supervised algorithm, similar to Espresso [14], for the automatic extraction of semantic relations from Wiktionary suggest that more and better results are obtained with the approach presented here, where manually created grammars based on lexical-syntactic patterns are used. On the one hand, in the machine learning approach the construction of the grammars, which is a time-consuming task, is not required. On the other hand, the grammar-based approach provides higher control over the patterns used. Furthermore, since the vocabulary used in dictionaries is typically simple, it should be possible to adapt grammars, originally created for the extraction of relations from a specific dictionary, to other dictionaries in the same language. This is what we have done for Wiktionary, using the grammars of PAPEL.

In a near future, the results of this research will be publicly available for the community and thus be a contribution for the development of better NLP applications for Portuguese.

Acknowledgements

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

References

1. Amsler, R.A.: A taxonomy for english nouns and verbs. In: Proceedings of 19th annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics, Morristown, NJ, USA (1981)
2. Calzolari, N., Pecchia, L., Zampolli, A.: Working on the italian machine dictionary: a semantic approach. In: Proceedings of 5th Conference on Computational Linguistics. pp. 49–52. Association for Computational Linguistics, Morristown, NJ, USA (1973)
3. Chodorow, M.S., Byrd, R.J., Heidorn, G.E.: Extracting semantic hierarchies from a large on-line dictionary. In: Proceedings of the 23rd annual meeting on Association for Computational Linguistics. pp. 299–304. Association for Computational Linguistics, Morristown, NJ, USA (1985)
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
5. Gonçalo Oliveira, H., Costa, H., Gomes, P.: Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In: Proceedings of INFORUM 2010 (September 2010)
6. 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)
7. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática 2(1), 77–93 (Maio 2010), nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009
8. Hearst, M.A.: Automated discovery of wordnet relations. In: [4], pp. 131–151 (1998)
9. Herbelot, A., Copestake, A.: Acquiring ontological relationships from wikipedia using RMRS. In: Proceedings of ISWC 2006 Workshop on Web Content Mining with Human Language Technologies (2006)
10. Ide, N., Veronis, J.: Knowledge extraction from machine-readable dictionaries: An evaluation. In: Steffens, P. (ed.) Machine Translation and the Lexicon, LNAI. Springer-Verlag (1995)
11. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from wikipedia. Intl. Journal of Human-Computer Studies (Maio 2009)
12. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, T.Y., Magistry, P., Huang, C.R.: Wiktionary and nlp: Improving synonymy networks. In: Proceedings Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. pp. 19–27. Association for Computational Linguistics, Suntec, Singapore (2009)
13. Nichols, E., Bond, F., Flickinger, D.: Robust ontology acquisition from machine-readable dictionaries. In: Kaelbling, L.P., Saffiotti, A. (eds.) Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI). pp. 1111–1116. Professional Book Center (2005)

14. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics. pp. 113–120. ACL Press, Sydney, Australia (2006)
15. Richardson, S.D., Dolan, W.B., Vanderwende, L.: Mindnet: Acquiring and structuring semantic information from text. In: Proceedings of 17th International Conference on Computational Linguistics (COLING). pp. 1098–1102 (1998)
16. Sagot, B., Fišer, D.: Building a Free French Wordnet from Multilingual Resources. In: Proceedings of LREC 2008 Workshop, OntoLex 2008. Marrackech, Morocco (2008)
17. Sajous, F., Navarro, E., Gaume, B., Prévot, L., Chudy, Y.: Semi-automatic endogenous enrichment of collaboratively constructed lexical resources: Piggybacking onto wiktionary. In: Advances in Natural Language Processing, 7th International Conference on NLP (IceTAL). LNCS, vol. 6233, pp. 332–344. Springer, Reykjavik, Iceland (2010)
18. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proc. 39th Annual Meeting of the Association for Computational Linguistics. pp. 442–449. ACL Press (9-11 July 2001)
19. Teixeira, J., Sarmento, L., Oliveira, E.: Comparing verb synonym resources for portuguese. In: Proceedings of Computational Processing of the Portuguese Language, 9th International Conference (PROPOR 2010). LNAI, vol. 6001, pp. 100–109. Springer (2010)
20. Wandmacher, T., Ovchinnikova, E., Krumnack, U., Dittmann, H.: Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In: Proceedings of 3rd Australasian Ontology Workshop (AOW 2007). CRPIT, vol. 85, pp. 61–69. ACS, Gold Coast, Australia (2007)
21. Weale, T., Brew, C., Fosler-Lussier, E.: Using the wiktionary graph structure for synonym detection. In: Proceedings of 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. pp. 28–31. People's Web '09, ACL Press, Stroudsburg, PA, USA (2009)
22. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of 6th International Language Resources and Evaluation (LREC'08). Marrakech, Morocco (2008)
23. Zesch, T., Müller, C., Gurevych, I.: Using wiktionary for computing semantic relatedness. In: Proceedings of 23rd National Conference on Artificial Intelligence (AAAI'08). pp. 861–866. AAAI Press (2008)