Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 45-52 recibido 18-11-2019 revisado 08-01-2020 aceptado 11-01-2020

Building wordnets with multi-word expressions from parallel corpora

Expansión de wordnets mediante unidades pluriverbales extraídas de corpus paralelos

Alberto Simões[1], **Xavier Gómez Guinovart**[2]
[1]2Ai – School of Technology, IPCA, Barcelos, Portugal
[2]Universidade de Vigo, SLI-TALG
asimoes@ipca.pt, xgg@uvigo.gal

Abstract: In this paper we present a method for enlarging wordnets focusing on multi-word terms and utilising data from parallel corpora. Our approach is validated using the Galician and Portuguese wordnets. The multi-word candidates obtained in this experiment were manually validated, obtaining a 73.2% accuracy for the Galician language and a 75.5% for the Portuguese language.
Keywords: wordnet, parallel corpora, lexical resources, multi-word expressions

Resumen: Presentamos un método para la ampliación de wordnets en el ámbito de las unidades pluriverbales, usando datos de corpus paralelos y aplicando el método a la expansión de los wordnets del gallego y del portugués. Las unidades pluriverbales que se obtienen en este experimento se validaron manualmente, obteniendo una precisión del 73.2% para el gallego y del 75.5% para el portugués.
Palabras clave: wordnet, corpus paralelos, recursos léxicos, unidades pluriverbales

1 Introduction

The Princeton WordNet (Miller et al., 1990)[1] (PWN) is, undoubtedly, a milestone in Natural Language Processing. This can be proven by the amount of wordnet-like projects available for most of the world languages. In this article we will discuss a methodology to enrich two different wordnets: Galnet (Gómez Guinovart and Solla Portela, 2018)[2] for Galician, and PULO (Simões and Gómez Guinovart, 2013)[3] for Portuguese.

Given the amount of manual work required to produce a quality resource, there have been different approaches to create new wordnets. Our proposal focuses only on the multi-word noun entries in wordnet, and on the usage of parallel corpora and translation patterns to obtain candidates for the target language.

This paper is structured as follows: Section 2 summarises the previous related work. Section 3 presents the resources used (wordnets, parallel corpora) and the derived resources (multi-word term list, annotated corpora, and probabilistic translation dictionaries). There follows Section 4 where the implemented algorithm is explained, and Section 5 where an evaluation of the obtained candidates is performed. Finally we draw some conclusions and present some directions for future work.

2 Related Work

Galnet and PULO wordnets have been created from the English WordNet 3.0, following the expand model (Vossen, 1998), where the variants associated with the PWN synsets are obtained through different strategies. The same approach has been taken in the MCR framework (González-Agirre and Rigau, 2013) for the creation of the wordnets of Spanish, Catalan and Basque.

One of the main methodologies used to extend a wordnet coverage from the variants associated with the PWN synsets is the acquisition of their translations from parallel corpora. Thus, in (Gómez Guinovart and Oliver, 2014) the authors apply that methodology to expand the Galnet first distribution from two different available parallel

[1] `https://wordnet.princeton.edu`
[2] `http://sli.uvigo.gal/galnet/`
[3] `http://wordnet.pt`

ISSN 1135-5948. DOI 10.26342/2020-64-5

textual resources: the automatically translated English–Galician SemCor Corpus;[4] and the English–Galician and Spanish–Galician sections of the CLUVI Corpus.[5] In either case, only the English or Spanish part of the parallel corpora has been sense-tagged for the experiment. In (Oliver, 2014) the same methodology is applied to the automatic translation of the English SemCor to six languages (Catalan, Spanish, French, German, Italian and Portuguese).

In (Simões and Gómez Guinovart, 2018), the authors used the Galician, Portuguese, Spanish, Catalan and English versions of the Bible from the CLUVI Corpus. They were annotated with part-of-speech and WordNet sense. The resulting synsets were aligned, and new variants for Galnet were extracted. After manual evaluation the approach presented a 96.8% accuracy. Unlike the research we present in this paper, all these previous experiments have been focused on monolexical extraction.

In (Gómez Guinovart and Simões, 2009) the authors presented a parallel corpora-based bilingual terminology extraction method based on the occurrence of bilingual morphosyntactic patterns in parallel text, with the support of probabilistic translation dictionaries for inter-language alignment. We applied this method using corpora for English–Galician and English–Portuguese, obtaining an accuracy rate between 87.4% and 96% depending on the characteristics of the corpus.

(Vintar and Fišer, 2008) present an approach to extend the automatically created Slovene wordnet with nominal multi-words from the English-Slovene part of the JRC-Acquis corpus of legislative text of the European Union by translating multi-words from Princeton WordNet with a technique that is based on word alignment and lexico-syntactic patterns. For each source multi-word, they extracted all sentence pairs from the parallel corpus that contain the source term. Also, for each single word from the source multi-word they extract all possible translation equivalents from the bilingual lexicon. Then, they use lexico-grammatical patterns to identify potential multi-word terms in the target language and check the word alignments for the selection of the best equivalent, which is the candidate with the most matches for each constituent word in the bilingual lexicon. The authors manually evaluated the set of candidate words obtained by this technique, filtered by a threshold of 0.05 as the lowest possible similarity score, obtaining an accuracy of 85% and a total of 1,059 new variants for the Slovene wordnet.

3 Resources

As pointed out before, the methodology we propose requires a source wordnet and a parallel corpus mapping the source wordnet language to the target language.

3.1 Wordnets

Both Galnet and PULO are part of the Multilingual Central Repository (MCR),[6] that currently integrates wordnets from six different languages (English, Spanish, Catalan, Galician, Basque and Portuguese) with WordNet 3.0 as Interlingual Index (ILI) (González-Agirre and Rigau, 2013). Table 1 provides the number of synsets and variants for the different languages gathered in this repository, and their percentage of development with respect to the English WordNet.

From the English WordNet a list of multi-word terms were extracted (Lloberes et al., 2013). There are 68,751 multi-word terms in the PWN. We decided to focus our experiment on the noun terms (63,073, above 90% of the total amount of multi-word terms). This list was then processed by FreeLing 4.1 (Padró and Stanilovsky, 2012)[7] in order to obtain each term's morphological structure, and understand which of these structures are more common and more likely to return interesting results.

3.2 Parallel corpora

Parallel corpora were obtained from the OPUS project.[8] For the English–Galician pair, we used the Gnome, KDE 4, Ubuntu, Tatoeba and OpenSubtitles2018 corpora, amounting to a total of 350,124 translation units. Given the limited existence of English–Galician corpora, there was no other reasonable alternative. For the English–Portuguese pair, only the OpenSubtitles2018 corpus was used, accounting for 26,805,614 translation units.

[4] http://gabormelli.com/RKB/SemCor_Corpus
[5] http://sli.uvigo.gal/CLUVI/
[6] http://adimen.si.ehu.es/web/MCR/
[7] http://nlp.lsi.upc.edu/freeling/
[8] http://opus.nlpl.eu

	English (PWN 3.0)		Galician (Galnet 3.0.28)	
	variants	**synsets**	**variants**	**synsets**
Total	206,941	117,659	70,056	43,057
%	100%	100%	33.8%	36.6%
	Spanish (MCR 2016)		**Portuguese (MCR 2016)**	
Total	146,501	78,995	32,604	17,942
%	70.8%	67.1%	15.8%	15.2%
	Catalan (MCR 2016)		**Basque (MCR 2016)**	
Total	100,793	60,956	50,037	30,263
%	48.7%	51.8%	24.2%	25.7%

Table 1: Current coverage of languages in MCR

These corpora were processed in two different ways:

- Lemmatisation and part-of-speech annotation using FreeLing. This annotation was performed without any kind of named entity or locution detection in the target language, so that the resulting corpus does not include any multi-word terms annotated. For the source language (English), the corpus was only tagged with the FreeLing multi-word recognition module using the list of multi-word expressions referred to in the previous section.
- Both the original corpus as the lemmatised corpus (result of the above annotation process) were subject of word-alignment using NATools (Simões and Almeida, 2003). Thus, for each language pair we obtained a probabilistic translation dictionary (PTD) for forms and lemmas. Each entry of a PTD maps a word in the source language to a set of probable translations as well as their translation probability.

4 Methodology

Our approach requires not just the resources presented in the previous section but also a set of morphosyntactic patterns. These patterns will later be used in the extraction algorithm to obtain variant candidates.

4.1 Morphosyntactic patterns

As pointed out before, this approach uses translation patterns: rules that make explicit the number of words (tokens) in each language, how the translation of each word switches its place during the translation process, and whether there is any addition or removal of words.

The rules were created starting from the list of multi-word terms present in the English WordNet, together with their morphological structure. Only the top 10 occurring structures were chosen for this experiment. A set of multi-word terms following each one of the morphological structures was compiled in order to manually study how their translation was performed. This study resulted in the patterns for English–Galician and for English–Portuguese presented in Figure 1, which cover about 90% of the multi-word nouns in WordNet. Note that FreeLing uses two different tag sets for English and for Portuguese/Galician, and that is why the prefixes presented on the left-hand side and the right-hand-side of the rule are different. Note also that both Galician and Portuguese share the same multi-word translation patterns for English–Galician and for English–Portuguese as a result of their similar morphological structures.

In the rules file, everything starting with a sharp (`#`) character is considered a comment. Then, each line is comprised of a left-hand side pattern, matching the English variant, and a right-hand side pattern that will try to match the corresponding Galician or Portuguese variant. Each item (token) in the pattern is separated from each other by a space. Each token can have a name (upper-case before the parenthesis) or a lemma (lower-case before the parenthesis), but never both, given that identifiers are used to match translations and, if a specific lemma is supplied, there is no translation check. Inside the parenthesis is the morphological category of the token being matched (using the beginning segment of any FreeLing tag).

```
#1. air compressor = compresor de aire / compressor de ar
A(NN) B(NN) =  B(NC) de(SP) o(DA)? A(NC)
#2. absolute zero = cero absoluto / zero absoluto
A(JJ) B(NN) =  B(NC) A(AQ)
#3. lubricating system = sistema de lubricación / sistema de lubrificação
A(VBG) B(NN) =  B(NC) de(SP) o(DA)? A(NC)
#4. analysis of variance = análise de varianza / análise de variância
A(NN) B(IN) C(NN) =  A(NC) B(SP) o(DA)? C(NC)
#5. closed circuit = circuíto pechado / circuito fechado
A(VBN) B(NN) = B(NC) A(VMP)
#6. communications satellite = satélite de comunicacións / satélite de comunicações
A(NNS) B(NN) =  B(NC) de(SP) o(DA)? A(NC)
#7. local area network = rede de área local / rede de área local
A(JJ) B(NN) C(NN) = C(NC) de(SP) o(DA)? B(NC) A(AQ)
#8. graphical user interface = interface gráfica de usuario / interface gráfica do utilizador
A(JJ) B(NN) C(NN) = C(NC) A(AQ) de(SP) o(DA)? B(NC)
#9. African green monkey = mono verde africano / macaco verde africano
A(JJ) B(JJ) C(NN) = C(NC) B(AQ) A(AQ)
#10. table of contents = táboa de contidos / tabela de conteúdo
A(NN) B(IN) C(NNS) = A(NC) B(SP) o(DA)? C(NC)
```

Figure 1: English–Galician/Portuguese rules and examples

As can be seen in the following specific pattern:

$$\text{A(NN) B(NN)} = \text{B(NC) de(SP) o(DA)? A(NC)}$$

the left-hand side is matching two nouns (tags starting with `NN`). The first one would be identified by A, and the second one by B. Therefore, considering the multi-word "*air compressor*" the following would be extracted:

$$\{A \mapsto \textit{air}, B \mapsto \textit{compressor}\}$$

The right-hand side of the rule specifies that the algorithm should look for a sequence of a noun (`NC`), a preposition (`SP`), an optional article (`DA`) and another noun. Note that the question mark following a token specifies that it is optional. Together with this sequence, the right-hand side also specifies that the first token found should be the translation of B, following by a token with lemma '*de*', another token with lemma '*o*', and finally a token that should be the translation of A.

4.2 Matching algorithm

The first part of the process is to create a reverse index. This index maps each nominal phrase from the list of multi-word terms to the translation units of the parallel corpora where they occur. By "translation unit" we refer to a pair of source and target sentences in the corpus, whether English–Galician or English–Portuguese.

Then, for each pair $(synset, variant)$ the following process is executed:

1. Ignore the pair if the source variant structure does not match any of the defined translation patterns.

2. For each translation pattern matching the source variant, search in the sequence of parts-of-speech for the target language if there is any occurrence for the right-hand side of the translation pattern.

3. If there is one or more occurrences of the target pattern, each one is evaluated, checking the probable translation probability with the source variant.

4. To evaluate the translation probability the rule placeholder identifiers come into play. The translation probability is computed for the words associated with the identifiers (for the source and target language) both for forms and lemmas. The same is done in the reverse order, as the probabilistic translation dictionaries are not symmetrical. For example, the pair (*air compressor*, *compresores de aire*) has $\{\text{A} \mapsto \textit{air}, \text{B} \mapsto \textit{compressor}\}$ for the source language, and $\{\text{A} \mapsto \textit{aire}, \text{B} \mapsto \textit{compressores}\}$ for the target language.

   Its translation probability in the source–

target direction is computed by:[9]

$$\begin{aligned}\mathcal{P} &= \frac{1}{4}\mathcal{P}_l\left(\mathcal{T}(air) = aire\right) \\ &+ \frac{1}{4}\mathcal{P}_f\left(\mathcal{T}(air) = aire\right) \\ &+ \frac{1}{4}\mathcal{P}_l\left(\mathcal{T}(compressor) = compresores\right) \\ &+ \frac{1}{4}\mathcal{P}_f\left(\mathcal{T}(compressor) = compresor\right)\end{aligned}$$

The same is done in the reverse direction, using the GL–EN and PT–EN dictionaries. The two probabilities obtained are then averaged.

5. Given all possible PoS alignments with the target sentence, only the one with greatest probability is kept.
6. Finally for all occurrences of the original variant, the target sequence whose alignment occurred more times is chosen.
7. In the final list of candidates, only the ones with probability greater than 0.1 were considered. This threshold was defined empirically.

5 *Evaluation and error analysis*

We have designed a protocol for the manual review of the extraction results by a lexicographer. Reviewing is done by evaluating the suitability of the candidates with respect to the WordNet sense taken into consideration. The evaluation of Portuguese results has been done in an exploratory mode without a preestablished error typology. After the elaboration of that typology, based on this previous evaluation of the Portuguese candidates, we have been able to register an error type for each bad candidate found during the evaluation of Galician results. Therefore, only in the case of Galician, erroneous candidates have also received a code that indicates the reason for their exclusion.

We have obtained 1,832 multi-word candidates for Galician and 12,172 for Portuguese. The reduced number of candidates, when compared with the total number of different multi-word expressions from English, is related to the corpora lexical variety. For instance, the English texts in the English–Galician corpus only contain 2,174 different multi-word expressions from the 60,073 included in the PWN list. At this moment, 500 candidates for Galician have been evaluated. The percentage of correct answers in the evaluated candidates reaches 73.2% of the cases. A similar number of candidates were evaluated for Portuguese, obtaining a 75.5% of correct answers in this case.

The difference in accuracy between the proposed approach and that of 85% reported in (Vintar and Fišer, 2008) can be attributed to two factors. On the one hand, Vintar and Fišer work with legal texts in the field of specialised terminology extraction, where accuracy tends to be higher than in general vocabulary acquisition. On the other hand, and although we cannot compare directly the score values as the similarity measures were computed by different algorithms, Vintar and Fišer use a threshold of 0.05 as the lowest possible similarity score, thus probably decreasing the coverage of their experiment.

At first our expectation was to have a high accuracy, given that multi-word terms are less ambiguous than single-word terms. Figures 2 and 3 show the ambiguity of single and multi-word terms. The X-axis is the number of synsets a variant belongs to, while the Y-axis is the number of variants. For example, there is a single-word variant belonging to 75 different synsets, while the most ambiguous multi-word variant just appears in 19 different synsets.

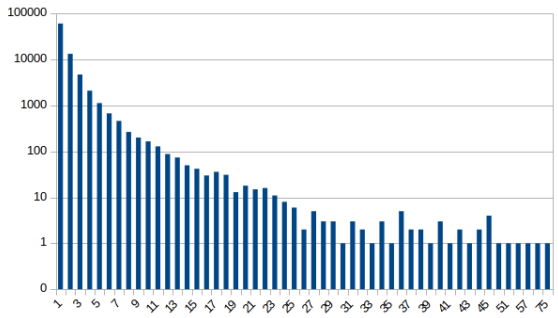

Figure 2: Ambiguity of single-word terms (number of variants vs number of synsets)

While these two images would confirm our expectation, some problems were found and erroneous candidates were generated for different reasons. In the following sections, we will describe and exemplify the most significant causes of error in the extraction process. Their incidence in the process of extracting Galician multi-word terms is shown in Table 2, where spelling errors are not considered, as explained below.

[9] $\mathcal{P}_f$ stands for the probability in the PTD computed from the forms corpus, while $\mathcal{P}_l$ is the probability from the lemmatised corpus.

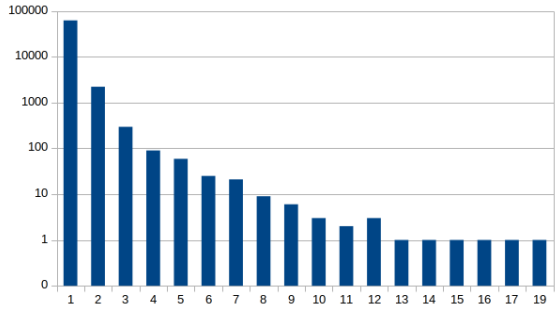

Figure 3: Ambiguity of multi-word terms (number of variants vs number of synsets)

Translation mistakes	18%
Idioms	26%
Transpositions	22%
Collocations	16%
Coordination	8%
Other types of error	10%

Table 2: Error typology in Galician multi-word term extraction

5.1 Spelling

In a few cases, the candidate generated from the corpus represents a variant rejected by the current official Galician normative. For example, the proposed candidate "*hospital siquiátrico*" for the concept of "*mental hospital*" (ili-30-03746574-n)[10] is not well written following the current regulations of the Galician language, which prescribe "*hospital psiquiátrico*" with initial "*p*" in the second word.

There are 8 errors of this type between the 500 candidates, from which 7 would be correct candidates with the corresponding spelling normativisation. These erroneous candidates cannot be considered as the result of any dysfunction of the extraction methodology, and can be easily identified and corrected during manual importation into the Galician wordnet, so they have not been taken into account for the evaluation of the accuracy of the results (and this is why they are not included in the data shown in Table 2).

5.2 Translation mistakes

Sometimes, the texts in the corpus contain translation mistakes that affect a multi-word in English and that lead to the generation of a wrong translation candidate.

For example, the English nominal compound "*numbers racket*", which has the meaning of "*an illegal daily lottery*" (ili-30-00508547-n),[11] is translated into Galician as "*raqueta de números*" (literally, "*racket of numbers*") in the OpenSubtitles2018 corpus, using the same words as in the source language, when the correct translation would be "*lotaría*" or "*lotaría ilegal*" depending on the context. In the same corpus, the English multi-word "*straight razor*" (ili-30-04332074-n)[12] has been rendered in Portuguese by "*gillete recta*" (literally, "*straight Gillette*"), when the correct translation would be "*navalha de barbear*", apart from the fact that Gillette is written with two t's. Therefore, these translation errors have led to the generation of the proposals "*gillete recta*" for Portuguese and "*raqueta de números*" for Galician which are bad translation candidates for their respective English terms.

5.3 Idioms

WordNet does not include free combinations, and some multi-word terms (like "*piece of cake*") have both an idiomatic sense (the one we found registered in WordNet) and the literal sense (not registered in WordNet).

Idioms are lexical sequences where the words mean something other than their literal meaning. In some cases, an erroneous candidate for an idiomatic multi-word is produced from its literal translation.

For instance, the erroneous proposal "*mina de sal*" for the meaning "*a job involving drudgery and confinement*" (ili-30-00606119-n)[13] is generated from the literal version of the English multiword "*salt mine*".

5.4 Transpositions

Shifts or transpositions in translation involve a change in the grammar from source language to target language. In some cases, the right translation from an English multi-word to a Galician or Portuguese term implies the change of the English noun group to a target language single noun.

This can cause errors in the application of the algorithm when it is possible to de-

[10] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-03746574-n

[11] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-00508547-n

[12] http://wordnet.pt/synset/04332074-n

[13] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-00606119-n

tect incorrect lexical alignments in the parallel corpora possibly due to a bad translation, as in the incorrect translation proposal for Galician "*pezas de mobiliario*" for the English "*piece of furniture*" (ili-30-00606119-n),[14] where the correct translation would be "*moble*"; or as in the improper candidate for Portuguese "*criança macho*" for the English "*male child*" (ili-30-10285313-n),[15] where the proper Portuguese equivalent would be "*menino*".

5.5 Collocations

In lexicology, collocations are sequences of two or more words that usually go together, like "*strong tea*" or "*false teeth*". Collocations are highly idiomatic and ruled by the norms of language use.

Sometimes extraction produces results that are grammatically and semantically correct in the target language, but do not follow its rules of use. For example, the translation proposed "*traballo de policía*" for "*police work*" (ili-30-00606119-n)[16] is incorrect in Galician, because usage prescribes "*traballo policial*" for this concept, with the adjective "*policial*" instead of the prepositional phrase "*de policía*". The same kind of error in extraction can be appreciated in the Portuguese candidate "*água santa*" for English "*holy water*" (ili-30-14846517-n),[17] where the language rules of use would prescribe "*água benta*" (literally, "*blessed water*").

5.6 Coordination

Extraction rules may fail when applied to coordinated structures. For example, when processing the EN–GL alignment from the OpenSubtitles2018 corpus:

> EN: *Iraq has chemical and biological weapons which could be activated within 45 minutes*
> GL: *Iraq posúe armas químicas e biolóxicas que poderían ser activadas en menos de 45 minutos*

the extraction algorithm proposes the equivalence between the original "*biological weapon*" and the translation "*armas químicas*" ("*chemical weapons*"), where the proper Galician equivalent would be "*armas biolóxicas*" ("*biological weapons*").

5.7 Other types of errors

Other types of errors occur with a lower level of significance. For instance, a possible cause of error, which occurs only twice for Galician in our evaluation, is the lexical ambiguity of the English multi-word in WordNet. In this case, the extraction process is applied to all the senses of the term, with a high risk of error. The lexical form "*sea horse*", for example, has two senses in the English WordNet: the first with the meaning of "*walrus*" (ili-30-02081571-n)[18] and the second with the meaning of "*small fish with horse like heads bent sharply downward and curled tails*" (ili-30-01456756-n)[19]. Because of this, the extraction algorithm proposes the incorrect equivalence between the English "*sea horse*" (ili-30-01456756-n) with the sense of small fish and the Galician "*morsa*" ("*walrus*"), where the proper Galician term would be "*cabaliño de mar*" (literally, "*little horse of the sea*").

6 Conclusions and future work

In this paper we proposed a methodology to extract multi-word variant candidates in Portuguese and Galician using the original multi-word variants from the English WordNet aided by parallel corpora and translation patterns. Despite the difficulties, the results of human evaluation in section 5 show that the presented methodology, applied to the enlargement of wordnets with general vocabulary, leads to results not so different from those reported in previous works in the field of specialised terminology extraction (Gómez Guinovart and Simões, 2009). This would demonstrate the importance of associating morphology-based translation patterns to lexical alignment for the identification of multi-word WordNet variant candidates in parallel corpora.

Although the obtained accuracy is reasonable, better results could be achieved using higher threshold values. Looking to the Portuguese language results, for instance, the average score for wrong variants is 0.2172, while for the correct variants is 0.2589. Even

[14] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-03405725-n

[15] http://wordnet.pt/synset/10285313-n

[16] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-00635012-n

[17] http://wordnet.pt/synset/14846517-n

[18] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-02081571-n

[19] http://sli.uvigo.gal/galnet/galnet_var.php?ili=ili-30-01456756-n

though the values are quite near, they show that it might be possible to obtain better accuracy values. Nevertheless, unlike (Vintar and Fišer, 2008), and given that all variants to be imported in Galnet and PULO will be manually validated, we preferred not to raise the threshold of the lowest possible similarity score and to obtain a bigger coverage of WordNet multi-word expressions.

The linguistic kinship between Portuguese and Galician has allowed us to apply the same techniques to carry out the task proposed in this research, and the results obtained for each language have been similar.

Future work includes both finishing the validation of the full set of Galician and Portuguese extracted variants, and introducing the validated variants in the Galnet and PULO knowledge databases.

7 Acknowledgements

This research has been carried out thanks to the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government and the European Fund for Regional Development (MCIU/AEI/FEDER), and was partially funded by Portuguese National funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

References

Gómez Guinovart, X. and A. Oliver. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50.

Gómez Guinovart, X. and A. Simões. 2009. Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns. In *Proc. of the Iberian SLTech 2009*, pages 13–16.

Gómez Guinovart, X. and M. A. Solla Portela. 2018. Building the Galician wordnet: Methods and applications. *Language Resources and Evaluation*, 52(1):317–339.

González-Agirre, A. and G. Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. *Linguamática*, 5(1):13–28.

Lloberes, M., A. Oliver, S. Climent, and I. Castellón. 2013. Tratamiento de multipalabras en WordNet 3.0. In A. L. et al., editor, *Applied Linguistics in the Age of Globalization*. Universitat de Lleida, Lleida, pages 141–158.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Oliver, A. 2014. WN-Toolkit: Automatic generation of wordnets following the expand model. In *Proc. of the 7th Global WordNet Conference*, pages 7–15, Tartu. GWN.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proc. of the Eight International Conference on Language Resources and Evaluation*, pages 2473–2479, Istanbul. ELRA.

Simões, A. and X. Gómez Guinovart. 2018. Extending the Galician wordnet using a multilingual Bible through lexical alignment and semantic annotation. volume 62 of *OpenAccess Series in Informatics*, pages 14:1–14:13, Dagstuhl. Schloss Dagstuhl.

Simões, A. and J. J. Almeida. 2003. NATools: A statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224.

Simões, A. and X. Gómez Guinovart. 2013. Dictionary alignment by rewrite-based entry translation. In J. P. Leal, R. Rocha, and A. Simões, editors, *2nd Symposium on Languages, Applications and Technologies*, volume 29 of *OpenAccess Series in Informatics*, pages 237–247, Dagstuhl. Schloss Dagstuhl.

Vintar, S. and D. Fišer. 2008. Harvesting multi-word expressions from parallel corpora. In *Proc. of the 6th International Conference on Language Resources and Evaluation*, pages 1091–1096, Marrakech. ELRA.

Vossen, P., editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer, Norwell.