WordNet-Shp: Towards the Building of a Lexical Database for a Peruvian Minority Language

Diego Maguiño-Valencia, Arturo Oncevay-Marcos and Marco A. Sobrevilla Cabezudo
Research Group on Artificial Intelligence (IA-PUCP)
Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú
{dmaguino,arturo.oncevay,msobrevilla}@pucp.edu.pe

Abstract

WordNet-like resources are lexical databases with highly relevance information and data which could be exploited in more complex computational linguistics research and applications. The building process requires manual and automatic tasks, that could be more arduous if the language is a minority one with fewer digital resources. This study focuses in the construction of an initial WordNet database for a low-resourced and indigenous language in Peru: Shipibo-Konibo (shp). First, the stages of development from a scarce scenario (a bilingual dictionary shp-es) are described. Then, it is proposed a synset alignment method by comparing the definition glosses in the dictionary (written in Spanish) with the content of a Spanish WordNet. In this sense, word2vec similarity was the chosen metric for the proximity measure. Finally, an evaluation process is performed for the synsets, using a manually annotated Gold Standard in Shipibo-Konibo. The obtained results are promising, and this resource is expected to serve well in further applications, such as word sense disambiguation and even machine translation in the shp-es language pair.

Keywords: WordNet, lexical database, minority language

1. Introduction

The building of digital linguistic resources is a great support for endangered languages, as they help to preserve relevant information and knowledge related, not only for the language itself, but also for the community whose speak it. Nevertheless, if those resources are not developed to be able for further analysis and research, they may be insufficient to assist in the preservation efforts (Berment, 2002).

In that context, computational linguistics is a research area that aims to understand linguistic phenomena, in an automatic way, through the processing and exploiting either linguistic corpora or language patterns from large amounts of data. In order to achieve that goal, the corpus must be in a machine-readable format, and might include structured information and linguistic meta-data that helps to automatically understand patterns from the language.

Among the most important lexical and structured resources, the WordNet is included (Fellbaum, 1998). This resource could be used as a thesaurus for different languages, and its exploitation might ease more complex tasks such as word sense disambiguation or even machine translation.

For that reason, this article describes the building of an initial version of a WordNet for an endangered language, which faces additional problems caused by the low-density of digital resources. The language case study is Shipibo-Konibo (SHP), one of the 47 indigenous languages spoken in Peru, specifically in the Amazonian region and has over 23,000 speakers. (Ministerio de Educación, Perú, 2013). Like most of it peers, SHP is classified as a minority language from both a social and a computational perspective (Forcada, 2006).

The paper is organized as follow. Section 2 defines shortly what a WordNet is. Next, Section 3 presents works regarding the main aspects in the building of WordNet-like resources for other languages. After that, the construction of the WordNet for Shipibo-Konibo is detailed in Section 4. Additionally, there is an evaluation process in Section 5. Finally, conclusions and future work are discussed.

2. WordNet

A WordNet is a lexical database in an specific language. WordNet contains words grouped into synonyms sets (synsets) where each synset represent a different sense and is identified by a interlingual index (ILI) that allows to work in multilingual contexts. In addition to having a synonyms set, each synset may show a definition (gloss) and also some use examples. The synsets are connected between them through semantic and lexical relationships like hypernym, hyponym, and others (Fellbaum, 1998).

There are shallow similarities between a WordNet and a thesaurus, which different sets of terms are grouped based on a meaning-similarity criteria. On the other hand, the labels of the Wordnet are defined by the semantic relationship between the words or entries, while the clusters of words in a synonym dictionary may not follow any distinctive pattern of explicit meaning similarity (Miller, 1995).

For the Spanish language, there are two main multilingual options that are vastly used: MultiWordNet (Emanuele et al., 2002) and Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012). Despite the fewer amount of synsets contained in the latter repository, MCR will be used in the study because its more recent updates.

Finally, a WordNet is established ideally as a free and open-source resource, so the goal is similar for the Shipibo-Konibo WordNet.

3. Related Works

The section describes the different aspects related to the development or improvement of WordNet-like resources for different languages, whether they are minority ones or not.

Farreres et al. (1998) presented a semi-automatic approach to address development of new WordNets, using the English version as a model. There is a manual alignment for

verbs and nouns, while the validation step was automated. The latter process chose the most relevant terms and develop a complete taxonomy of semantic primitives. The proposed method was applied to build both Spanish and Catalan WordNets, using bilingual dictionaries and lexicons as the main sources.

Semi-automatic procedures were also proposed for building WordNets in other languages. For Persian, there were special considerations regarding the verb composition, as they are formed by more than 2 verbs usually (Rouhizadeh et al., 2008). Besides, an Ancient Greek WordNet was developed from a Greek-English dictionary, by supposing a semantic closeness regarding the terms translated (Bizzoni et al., 2014).

Thai WordNet was built using an own system called WordNet Builder (Sathapornrungkij and Pluempitiwiriyawej, 2005). The WordNet in English and machine readable dictionaries (MRD) were the main input sources, and both were connected trough a Link Analyzer for synset match. The validation process was performed by a statistical classifier, reducing human intervention.

In the Polish version, the validation process was performed with a similarity measure with the English WordNet, enhanced by human evaluation later (Broda et al., 2008). For the study, the synsets were built using a metric for semantic relation between different terms and POS-tags.

There were other studies focused in the improvements of previous WordNets. Mititelu (2012) proposed an addition of morpho-semantic relations for the Romanian WordNet. Heuristics were applied to group the terms and affixes in pairs. Then, these pairs were semantically tagged in three levels: monolingual, multilingual and for different NLP applications. Likewise, Bond et al. (2009) addressed improvements in the Japanese WordNet by increasing the vocabulary coverage, annotating more bilingual English-Japanese texts, and connecting the WordNet with different resources such as lexicons or image repositories.

Finally, Taghizadeh and Faili (2016) focused their efforts in building WordNet for low-resource scenarios. Using an Expectation-Maximization (EM) algorithm plus a cross-lingual word sense disambiguation method, a high quality WordNet could be built for Persian. Other positive aspect was the use of minimal resources, such as a bilingual dictionary and a monolingual textual corpus.

4. WordNet-Shp

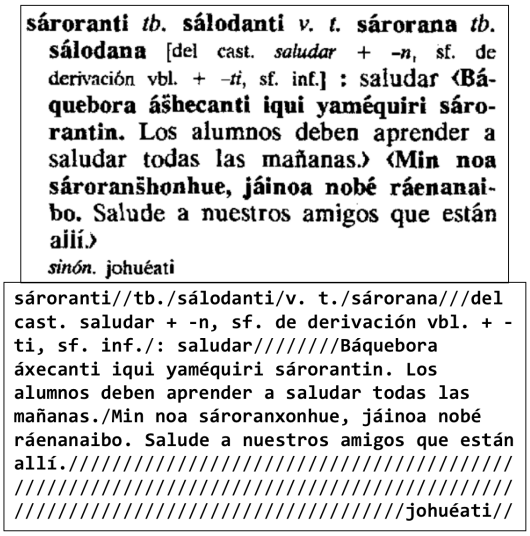
This section includes the steps followed for the development of the initial WordNet-Shp, a WordNet-like resource for Shipibo-Konibo. There are two main phases. The first phase consisted in the digitalization and pre-processing of a bilingual dictionary shp-es (Spanish). The second one consisted on the synset alignment task by using a similarity metric (provided by word2vec) with the definition glosses in the dictionary and the Spanish WordNet of Multilingual Central Repository (MCR)(Gonzalez-Agirre et al., 2012). The words of different glosses are compared against each other to obtain an average for each word obtained from the gloss of the dictionary and then averaged to obtain a final result the synset. The highest result is chosen.

The processed dictionary, the aligned synsets, and the Gold Standard for evaluation which consisted of a hundred synsets (see Subsection 5.1) are available in a project site[1].

4.1. Pre-Processing a Dictionary

An algorithm was built to automatically extract the words from an old-fashioned Spanish-Shipibo bilingual dictionary (Lauriout et al., 1993). For each word entry, an structured output is obtained, which includes several fields.First, there are 9 unique fields: term, reference, type of variant of the term, variant, grammatical category (POS-tag), main part, type of variant of the main part, variant of the main part, etymology. Then, given that a term might have more than one sense, each sense must be stored separately. So it was confirmed that the maximum number of different senses per word in the dictionary was 12. Each sense includes 6 glosses, 2 usage notes and 3 examples. Thus, it makes a total of eleven fields for every sense. Finally, there are 3 unique fields at the end of the entry: Synonymous paragraph, parent term, sub-entry type. The latter two only applies when the term is a sub-entry and is related to a parent main entry.

For instance, the parsed output for the term *sároranti* would be structured as in Figure 1.



sároranti *tb.* **sálodanti** *v. t.* **sárorana** *tb.* **sálodana** [del cast. *saludar* + *-n*, sf. de derivación vbl. + *-ti*, sf. inf.] : saludar ‹**Báquebora áshecanti iqui yaméquiri sárorantin.** Los alumnos deben aprender a saludar todas las mañanas.› ‹**Min noa sároranshonhue, jáinoa nobé ráenanaibo.** Salude a nuestros amigos que están allí.›

sinón. johuéati

```
sároranti//tb./sálodanti/v. t./sárorana///del
cast. saludar + -n, sf. de derivación vbl. + -
ti, sf. inf./: saludar////////Báquebora
áxecanti iqui yaméquiri sárorantin. Los
alumnos deben aprender a saludar todas las
mañanas./Min noa sároranxonhue, jáinoa nobé
ráenanaibo. Salude a nuestros amigos que están
allí.////////////////////////////////////////
////////////////////////////////////////////
/////////////////////////////////johuéati//
```

Figure 1: (Top) Original dictionary entry for *sároranti*. (Bottom) Parsed output for the entry. There are many separators ('/') together due to the possibility of extracting other elements between them. In this case, there is only one word sense, so there are a lot of separators together at the end, since up to 12 senses are expected as maximum.

The total amount of entries stored is about 5800, including 4815 terms from the main grammatical word classes (nouns, verbs, adjectives and adverbs). The remaining 985 are suffixes, prefixes, conjunctions, prepositions, among other categories. A distribution of the main word entries and their respective amount of senses in the Shipibo-Konibo Wordnet is presented in Table 1.

[1]WordNet-Shp data available in: chana.inf.pucp.edu.pe/resources/wordnet-shp

POS / #s.	Nouns	Verbs	Adj.	Adv.
1	2 231	1 453	357	96
2	174	266	62	11
3	59	48	13	2
4	14	16	1	0
5	6	3	0	0
6	2	0	0	0
7	0	0	0	1
Total	2 486	1 786	433	110

Table 1: Distribution of words with and without ambiguous senses found in the Shipibo-Konibo dictionary: Number of senses (#s.) per Part-of-Speech (POS) tag

4.2. Synsets Alignment

The alignment algorithm focuses in comparing the glosses of the word entries in Shipibo with the data of a Spanish WordNet, in order to obtain the closeness between the meanings among them. To define which synset each word sense belongs to, all the synsets in Spanish (terms, glosses and examples) were taken into account.
For this research, a word2vec model (Mikolov et al., 2013) was trained based on a general corpus (without annotation) of the Spanish language composed by approximately 1.4 billion words (Cardellino, 2016). The corpus was created by compiling various resources of the Spanish language that can be found on the Internet. Some of them are listed below:

- Collection of legal texts in Spanish from the European Union
- United Nations documents
- Parts of corpus in Spanish in other languages
- Protected articles from the Wikipedia in Spanish

With the trained word2Vec model, the classification algorithm for the synset alignment was the next step. The pseudo-code is presented below:

Algorithm 1 Synset alignment

for $i = 1$ to $|V|$ **do**
  for $j = 1$ to $|s_{v_i}|$ **do**
    $max_{ij} = 0$;
    for $k = 1$ to $|W_{es}|$ **do**
      $sim_{ijk} = word2vec\_similarity(s_{jv_i}, W_{es_k})$
      if $sim_{ijk} > max_{ij}$ **then**
        $max_{ij} = sim_{ijk}$
      end if
    end for
    $insert\_to\_BD(s_{jv_i}, W_{es_{argmax(k)}})$
  end for
end for

Where $|V|$ is the size of vocabulary V, $|s_{v_i}|$ is the number of different senses for the word s_{v_i}, and $|W_{es}|$ is the size of entries in the Spanish WordNet (W_{es}). word2vec_similarity is obtained by calculating the cosine between the two vectors.
Regarding the pseudocode, this algorithm works as follows: Each sense of each word found in the Shipibo dictionary was compared to each synset of the WordNet in Spanish using the word2vec model previously trained to obtain a similarity metric. This measurement was expressed as a decimal number.
For the calculation, it was taken the same word, glosses and corresponding examples, if any, of each Spanish word found in the Shipibo gloss. The glosses were filtered to consider only the words of the categories that arise more frequently and that allow to understand the context. These categories are nouns and verbs; and to a lesser extent, adjectives and adverbs.
For example, in the case of a noun, the nouns and verbs of its gloss were considered (due to the strong connection between the two categories). Each word was taken only once, eliminating repetitions because it could be the case that the terms or words contained in the glosses would be repeated.
Additionally, the Tree-tagger (Schmid, 1995) was used to extract grammatical categories and lemmas from words in Spanish. For instance, for the verb *manéxti* (”clean” in English), which was taken in the sense whose gloss is ”to tie the hairs of the head or crown with other hairs of the same head”, the process would be as follows:

- Tie the hairs of the head or the crown with other hairs his own head to
- Tying, hair, head, crown, hair, head - Only verbs and nouns
- Tie, hair, head, crown - Lemmas were extracted and word repetitions were erased

The similarity was expressed as a decimal number between -1 and 1 (being the result of a cosine). The greater the result, the relationship between words would be closer. Therefore, to decide which synset corresponds to the word in Shipibo, the synset with the greater similarity is considered. Each time the synset corresponding to a word in Shipibo was found, everything related to the term was inserted in the database following the standard used by MCR (Gonzalez-Agirre et al., 2012).
Once the algorithm with two words was executed in Shipibo as a sample, the nearest Spanish synset was obtained, and that was found by the classification algorithm described. For example, *mocoxoti* (which means ”clean”) was correctly classified into one of the synsets where the word ”clean” (in Spanish) is found in the same sense as to sort or arrange. It also shows all the words associated with the synset, as well as its ili code (used as the international standard).

5. Evaluation

To carry out the evaluation a gold standard was needed. It was prepared manually by a group of linguists and native speakers of Shipibo. The evaluation metric in this study is the accuracy, which is calculated after testing the classification algorithm with the gold standard.

Category	# Synsets
Nouns	105
Verbs	43
Adjetives	8
Adverbs	2

Table 2: Current state of the WordNet-Shp

5.1. Gold Standard

The gold standard was made by linguists and a native Shipibo-Konibo speaker. They selected a group of words extracted from the WordNet in Spanish and put all the possible synonyms for each one. In this way, a hundred synsets were formed manually, separated by grammatical categories as thus: 76 formed by nouns, 15 by verbs, 7 by adjectives and 2 by adverbs.

5.2. Results

All the words of the gold standard were evaluated to analyze if the retrieved synset is really the corresponding one. The total accuracy obtained with the gold standard was 32.8%. Some samples of classified synsets are presented below:

- Synset: spa-30-03571439-n Word: *chachí* (injector). Gloss: injector. Synset Result: spa-30-03571439-n
- Synset: spa-30-05599617-n Word: *cói* (chin). Gloss: protruding part of the jaw. Synset Result: spa-30-05598147-n
- Synset: spa-30-03343853-n Word: *tóoati,tsakati* (gun) Gloss: portable weapon. Synset Result: spa-30-00001740-n

In the first sample, there is a match because the original and the retrieved synset are the same. In the second example, the synsets do not match so the classification was labeled as incorrect. However, in the synset result (spa-30-05598147-n) the word "nose" shows up, which is a part of one's face so it's related to what we were originally looking for (chin). In the third and final example, the classification was not correct either but there is a lexical gap because the word *tsakati* wasn't found in the dictionary as a entry.
The precision obtained might be due to the following points:

- The obtained synset makes sense but it's not the same ILI. This could happen because the wordnet is very refined
- Some or several words used in the gloss of a particular word do not bear much relation with it
- Minor errors in the dictionary processing

Finally, after processing words from the dictionary that were included in the gold standard and an extra hundred (200 in total), the current state of the Wordnet in Shipibo by number of synsets is presented in Table 2.

6. Conclusions and Future Work

This study aimed the development of a new lexical and synonym-based resource for the Shipibo-Konibo language. Likewise, this repository uses the international standard for locating translations in other languages based on the synsets codification. For this purpose, a bilingual dictionary was pre-processed for extracting information of word entries and their senses in Shipibo-Konibo. After that, using the Spanish WordNet, an algorithm aligned each word sense in Shipibo-Konibo with its Spanish peer. The existing relationships in the Spanish WordNet were considered in order to be inherited in the Shipibo-Konibo repository.
Regarding the evaluation, there was a manual analysis of the synset quality. This was supported by professional linguists, and the output was the development of the Gold Standard. The results of the alignment showed a close similarity in the word sense distribution between Shipibo-Konibo and Spanish. This is caused mainly by the high presence of unique-sense words. Besides, as the number of senses is higher, the amount of words decreases considerably.
Finally, the developed resource stores all the words including their different senses in Shipibo-Konibo. Also, there is a web interface under development for querying the entries. All of these resources will be available in the following link: https://github.com/iapucp/wordnet-shp-lrec2018
As future work, we want to improve our algorithm of synset alignment and to include other relations between synsets, like hypernyms or hyponyms. We believe that it will not be difficult because it is possible to bring these relations from WordNets in other languages (like Spanish).

Acknowledgments

We highly appreciate the linguistic team effort that made possible the creation of this resource: Dr. Roberto Zariquiey, Alonso Vásquez, Gabriela Tello, Renzo Ego-Aguirre, Lea Reinhardt and Marcela Castro. We are also thankful to our native speakers (Shipibo-Konibo) collaborators: Juan Agustín, Carlos Guimaraes, Ronald Suárez and Miguel Gomez. Finally, we gratefully acknowledge the support of the "Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC, Peru) under the contract 225-2015-FONDECYT.

7. Bibliographical References

Berment, V. (2002). Several directions for minority languages computerization. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–5. Association for Computational Linguistics.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. R. (2014). The making of Ancient Greek WordNet. In *LREC*, volume 2014, pages 1140–1147.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., and Kanzaki, K. (2009). Enhancing the japanese wordnet. In *Proceedings of the 7th workshop on Asian language resources*, pages 1–8. Association for Computational Linguistics.

Broda, B., Derwojedowa, M., Piasecki, M., and Szpakowicz, S. (2008). Corpus-based semantic relatedness for the construction of Polish WordNet. In *LREC*.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings, March.

Emanuele, P., Luisa, B., and Christian, G. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet, Mysore, India*.

Farreres, X., Rigau, G., and Rodriguez, H. (1998). Using wordnet for building wordnets. *arXiv preprint cmp-lg/9806016*.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Forcada, M. (2006). Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC*, volume 6, pages 1–6.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Lauriout, E., Day, D., and Loriot, J. (1993). Diccionario Shipibo-Castellano.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ministerio de Educación, Perú. (2013). *Documento nacional de lenguas originarias del Perú*. MINEDU. URI: http://repositorio.minedu.gob.pe/handle/123456789/3549.

Mititelu, V. B. (2012). Adding morpho-semantic relations to the Romanian Wordnet. In *LREC*, pages 2596–2601.

Rouhizadeh, M., Shamsfard, M., and Yarmohammadi, M. A. (2008). Building a WordNet for Persian verbs. In *in the Proceedings of the Fourth Global WordNet Conference (GWC'08). The Fourth Global WordNet Conference, 2008*. Citeseer.

Sathapornrungkij, P. and Pluempitiwiriyawej, C. (2005). Construction of Thai WordNet lexical database from machine readable dictionaries. *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Schmid, H. (1995). Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Taghizadeh, N. and Faili, H. (2016). Automatic Wordnet development for low-resource languages using cross-lingual WSD. *J. Artif. Intell. Res.(JAIR)*, 56:61–87.