Automatically constructing Wordnet synsets

Khang Nhut Lam, Feras Al Tarouti and Jugal Kalita
Computer Science department
University of Colorado
1420 Austin Bluffs Pkwy, Colorado Springs, CO 80918, USA
{klam2,faltarou,jkalita}@uccs.edu

Abstract

Manually constructing a Wordnet is a difficult task, needing years of experts' time. As a first step to automatically construct full Wordnets, we propose approaches to generate Wordnet synsets for languages both resource-rich and resource-poor, using publicly available Wordnets, a machine translator and/or a single bilingual dictionary. Our algorithms translate synsets of existing Wordnets to a target language *T*, then apply a ranking method on the translation candidates to find best translations in *T*. Our approaches are applicable to any language which has at least one existing bilingual dictionary translating from English to it.

1 Introduction

Wordnets are intricate and substantive repositories of lexical knowledge and have become important resources for computational processing of natural languages and for information retrieval. Good quality Wordnets are available only for a few "resource-rich" languages such as English and Japanese. Published approaches to automatically build new Wordnets are manual or semi-automatic and can be used only for languages that already possess some lexical resources.

The Princeton Wordnet (PWN) (Fellbaum, 1998) was painstakingly constructed manually over many decades. Wordnets, except the PWN, have been usually constructed by one of two approaches. The first approach translates the PWN to *T* (Bilgin et al., 2004), (Barbu and Mititelu, 2005), (Kaji and Watanabe, 2006), (Sagot and Fišer, 2008), (Saveski and Trajkovsk, 2010) and (Oliver and Climent, 2012); while the second approach builds a Wordnet in *T*, and then aligns it with the PWN by generating translations (Gunawan and Saputra, 2010). In terms of popularity, the first approach dominates over the second approach. Wordnets generated using the second approach have different structures from the PWN; however, the complex agglutinative morphology, culture specific meanings and usages of words and phrases of target languages can be maintained. In contrast, Wordnets created using the first approach have the same structure as the PWN.

One of our goals is to automatically generate high quality synsets, each of which is a set of cognitive synonyms, for Wordnets having the same structure as the PWN in several languages. Therefore, we use the first approach to construct Wordnets. This paper discusses the first step of a project to automatically build core Wordnets for languages with low amounts of resources (viz., Arabic and Vietnamese), resource-poor languages (viz., Assamese) or endangered languages (viz., Dimasa and Karbi)[1]. The sizes and the qualities of freely existing resources, if any, for these languages vary, but are not usually high. Hence, our second goal is to use a limited number of freely available resources in the target languages as input to our algorithms to ensure that our methods can be felicitously used with languages that lack much resource. In addition, our approaches need to have a capability to reduce noise coming from the existing resources that we use. For translation, we use a free machine translator (MT) and restrict ourselves to using it as the only "dictionary" we can have. For research purposes, we have obtained free access to the Microsoft Translator, which supports translations among 44 languages. In particular, given public Wordnets aligned to the PWN (such as the FinnWordNet (FWN) (Lindén, 2010) and the JapaneseWordNet (JWN) (Isahara et al., 2008)) and the Microsoft Translator, we build Wordnet synsets for *arb, asm, dis, ajz* and *vie*.

[1]ISO 693-3 codes of Arabic, Assamese, Dimasa, Karbi and Vietnamese are *arb, asm, dis, ajz* and *vie*, respectively.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 106–111,
Baltimore, Maryland, USA, June 23-25 2014.

2 Proposed approaches

In this section, we propose approaches to create Wordnet synsets for a target languages *T* using existing Wordnets and the MT and/or a single bilingual dictionary. We take advantage of the fact that every synset in PWN has a unique *offset-POS*, referring to the offset for a synset with a particular part-of-speech (POS) from the beginning of its data file. Each synset may have one or more words, each of which may be in one or more synsets. Words in a synset have the same sense. The basic idea is to extract corresponding synsets for each *offset-POS* from existing Wordnets linked to PWN, in several languages. Next, we translate extracted synsets in each language to *T* to produce so-called *synset candidates* using MT. Then, we apply a ranking method on these candidates to find the correct words for a specific *offset-POS* in *T*.

2.1 Generating synset candidates

We propose three approaches to generate synset candidates for each *offset-POS* in *T*.

2.1.1 The direct translation (DR) approach

The first approach directly translates synsets in PWN to *T* as in Figure 1.

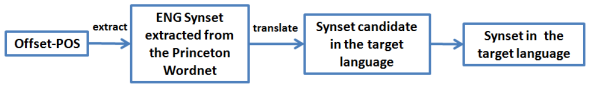

Figure 1: The DR approach to construct Wordnet synsets in a target language *T*.

For each *offset-POS*, we extract words in that synset from the PWN and translate them to the target language to generate translation candidates.

2.1.2 Approach using intermediate Wordnets (IW)

To handle ambiguities in synset translation, we propose the IW approach as in Figure 2. Publicly available Wordnets in various languages, which we call intermediate Wordnets, are used as resources to create synsets for Wordnets. For each *offset-POS*, we extract its corresponding synsets from intermediate Wordnets. Then, the extracted synsets, which are in different languages, are translated to *T* using MT to generate synset candidates. Depending on which Wordnets are used and the number of intermediate Wordnets, the number of candidates in each synset and the number of synsets in the new Wordnets change.

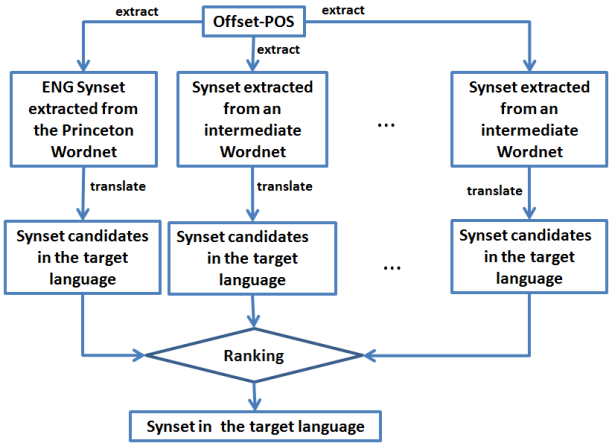

Figure 2: The IW approach to construct Wordnet synsets in a target language *T*

2.1.3 Approach using intermediate Wordnets and a dictionary (IWND)

The IW approach for creating Wordnet synsets decreases ambiguities in translations. However, we need more than one bilingual dictionary from each intermediate languages to *T*. Such dictionaries are not always available for many languages, especially the ones that are resource poor. The IWND approach is like the IW approach, but instead of translating immediately from the intermediate languages to the target language, we translate synsets extracted from intermediate Wordnets to English (*eng*), then translate them to the target language. The IWND approach is presented in Figure 3.

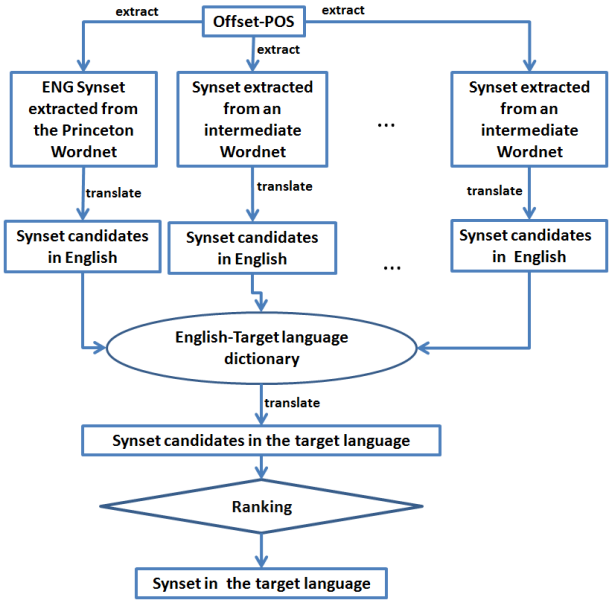

Figure 3: The IWND approach to construct Wordnet synsets

2.2 Ranking method

For each of *offset-POS*, we have many translation candidates. A translation candidate with a higher rank is more likely to become a word belonging to the corresponding *offset-POS* of the new Wordnet in the target language. Candidates having the same ranks are treated similarly. The rank value in the range 0.00 to 1.00. The rank of a word w, the so-called $rank_w$, is computed as below.

$$rank_w = \frac{occur_w}{numCandidates} * \frac{numDstWordnets}{numWordnets}$$

where:

- *numCandidates* is the total number of translation candidates of an *offset-POS*
- $occur_w$ is the occurrence count of the word w in the *numCandidates*
- *numWordnets* is the number of intermediate Wordnets used, and
- *numDstWordnets* is the number of distinct intermediate Wordnets that have words translated to the word w in the target language.

Our motivation for this rank formula is the following. If a candidate has a higher occurrence count, it has a greater chance to become a correct translation. Therefore, the occurrence count of each candidate needs to be taken into account. We normalize the occurrence count of a word by dividing it by *numCandidates*. In addition, if a candidate is translated from different words having the same sense in different languages, this candidate is more likely to be a correct translation. Hence, we multiply the first fraction by *numDstWordnets*. To normalize, we divide results by the number of intermediate Wordnet used.

For instance, in our experiments we use 4 intermediate Wordnets, viz., PWN, FWN, JWN and WOLF Wordnet (WWN) (Sagot and Fišer, 2008). The words in the *offset-POS* "00006802-v" obtained from all 4 Wordnets, their translations to *arb*, the occurrence count and the rank of each translation are presented in the second, the fourth and the fifth columns, respectively, of Figure 4.

2.3 Selecting candidates based on ranks

We separate candidates based on three cases as below.

Case 1: A candidate w has the highest chance to become a correct word belonging to a specific synset in the target language if its rank is 1.0. This means that all intermediate Wordnets contain the synset having a specific *offset-POS* and all words belonging to these synsets are translated to the same word w. The more the number of intermediate Wordnets used, the higher the chance the candidate with the rank of 1.0 has to become the correct translation. Therefore, we accept all translations that satisfy this criterion. An example of this scenario is presented in Figure 5.

Words	Cand.	TL	Occur	Rank
chuff^A	شوف	shwf	1	0.036
huff^A	هوف	hwf	1	0.036
puff^A	نفخة	nfkhh	2	0.143
puuskutta^B	بووسكوتا	bwwskwta	1	0.036
puhkua^B	بوهكوا	bwhkwa	1	0.036
läähättää^B	أنفاسها	anfasoha	1	0.036
bouffée^C	نفخة	nfkhh	2	0.143

Figure 4: Example of calculating the ranks of candidates translated from words belonging to the *offset-POS* "00006802-v" in 4 Wordnets: PWN, FWN, JWN and WWN. The $word^A$, $word^B$ and $word^C$ are obtained from PWN, FWN and WWN, respectively. The JWN does not contain this *offset-POS*. *TL* presents transliterations of the words in *arb*. The *numWordnets* is 4 and the *numCandidates* is 7. The rank of each candidate is shown in the last column of Figure 4.

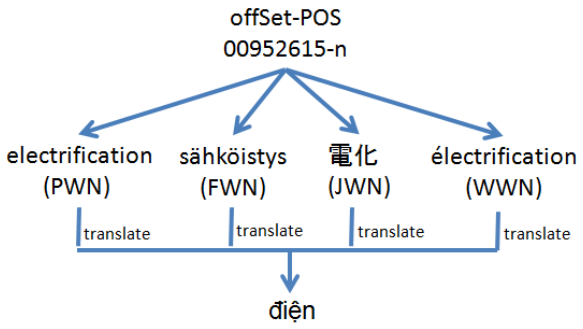

Figure 5: Example of Case 1: Using the IW approach with four intermediate Wordnets, PWN, FWN, JWN and WWN. All words belonging to the *offSet-POS* "00952615-n" in all 4 Wordnets are translated to the same word "điện" in *vie*. The word "điện" is accepted as the correct word belonging to the *offSet-POS* "00952615-n" in the Vietnamese Wordnet we create.

Case 2: If an *offSet-POS* does not have candidates having the rank of 1.0, we accept the candidates having the greatest rank. Figure 6 shows the example of the second scenario.

Case 3: If all candidates of an *offSet-POS* has the same rank which is also the greatest rank, we

Wordnet	Words	Cand.	Rank
PWN	send	gửi	0.67
PWN	send out	gửi	0.67
FWN	lähettää	gửi	0.67
WWN	transmettre	truyền tải	0.06
WWN	virer	chuyển giao	0.06
WWN	envoyer	gửi	0.67

Figure 6: Example of Case 2: Using the IW approach with three intermediate Wordnets, PWN, FWN and WWN. For the *offSet-POS* "01437254-v", there is no candidate with the rank of 1.0. The highest rank of the candidates in "vie" is 0.67 which is the word gửi. We accept "gửi" as the correct word in the *offSet-POS* "01437254-v" in the Vietnamese Wordnet we create.

skip these candidates. Table 1 gives an example of the last scenario.

Wordnet	Words	Cand.	Rank
PWN	act	hành động	0.33
PWN	behave	hoạt động	0.33
FWN	do	làm	0.33

Table 1: Example of Case 3: Using the DR approach. For the *offSet-POS* "00010435-v", there is no candidate with the rank of 1.0. The highest rank of the candidates in *vie* is 0.33. All of 3 candidates have the rank as same as the highest rank. Therefore, we do not accept any candidate as the correct word in the *offSet-POS* "00010435-v" in the Vietnamese Wordnet we create.

3 Experiments

3.1 Publicly available Wordnets

The PWN is the oldest and the biggest available Wordnet. It is also free. Wordnets in many languages are being constructed and developed[2]. However, only a few of these Wordnets are of high quality and free for downloading. The EuroWordnet (Vossen, 1998) is a multilingual database with Wordnets in European languages (e.g., Dutch, Italian and Spanish). The AsianWordnet[3] provides a platform for building and sharing Wordnets for Asian languages (e.g., Mongolian, Thai and Vietnamese). Unfortunately, the progress in building most of these Wordnets is slow and they are far from being finished.

[2]http://www.globalWordnet.org/gwa/Wordnet_table.html
[3]http://www.asianWordnet.org/progress

In our current experiments as mentioned earlier, we use the PWN and other Wordnets linked to the PWN 3.0 provided by the Open Multilingual Wordnet[4] project (Bond and Foster, 2013): WWN, FWN and JWN. Table 2 provides some details of the Wordnets used.

Wordnet	**Synsets**	**Core**
JWN	57,179	95%
FWN	116,763	100%
PWN	117,659	100%
WWN	59,091	92%

Table 2: The number of synsets in the Wordnets linked to the PWN 3.0 are obtained from the Open Multilingual Wordnet, along with the percentage of synsets covered from the semi-automatically compiled list of 5,000 "core" word senses in PWN. Note that synsets which are not linked to the PWN are not taken into account.

For languages not supported by MT, we use three additional bilingual dictionaries: two dictionaries *Dict(eng,ajz)* and *Dict(eng,dis)* provided by Xobdo[5]; one *Dict(eng,asm)* created by integrating two dictionaries *Dict(eng,asm)* provided by Xobdo and Panlex[6]. The dictionaries are of varying qualities and sizes. The total number of entries in *Dict(eng,ajz)*, *Dict(eng,asm)* and *Dict(eng,dis)* are 4682, 76634 and 6628, respectively.

3.2 Experimental results and discussion

As previously mentioned, our primary goal is to build high quality synsets for Wordnets in languages with low amount of resources: *ajz, asm, arb, dis* and *vie*. The number of Wordnet synsets we create for *arb* and *vie* using the DR approach and the coverage percentage compared to the PWN synsets are 4813 (4.10%) and 2983 (2.54%), respectively. The number of synsets for each Wordnet we create using the IW approach with different numbers of intermediate Wordnets and the coverage percentage compared to the PWN synsets are presented in Table 3.

For the IWND approach, we use all 4 Wordnets as intermediate resources. The number of Wordnet synsets we create using the IWND approach are presented in Table 4. We only construct Wordnet synsets for *ajz*, *asm* and *dis* using the IWND ap-

[4]http://compling.hss.ntu.edu.sg/omw/
[5]http://www.xobdo.org/
[6]http://panlex.org/

App.	Lang.	WNs	Synsets	% coverage
IW	arb	2	48,245	41.00%
IW	vie	2	42,938	36.49%
IW	arb	3	61,354	52.15%
IW	vie	3	57,439	48.82%
IW	arb	4	75,234	63.94%
IW	vie	4	72,010	61.20%

Table 3: The number of Wordnet synsets we create using the IW approach. *WNs* is the number of intermediate Wordnets used: 2: PWN and FWN, 3: PWN, FWN and JWN and 4: PWN, FWN, JWN and WWN.

proach because these languages are not supported by MT.

App.	Lang.	Synsets	% coverage
IWND	ajz	21,882	18.60%
IWND	arb	70,536	59.95%
IWND	asm	43,479	36.95%
IWND	dis	24,131	20.51%
IWND	vie	42,592	36.20%

Table 4: The number of Wordnets synsets we create using the IWND approach.

Finally, we combine all of the Wordnet synsets we create using different approaches to generate the final Wordnet synsets. Table 5 presents the final number of Wordnet synsets we create and their coverage percentage.

Lang.	Synsets	% coverage
ajz	21,882	18.60%
arb	76,322	64.87%
asm	43,479	36.95%
dis	24,131	20.51%
vie	98,210	83.47%

Table 5: The number and the average score of Wordnets synsets we create.

Evaluations were performed by volunteers who use the language of the Wordnet as mother tongue. To achieve reliable judgment, we use the same set of 500 *offSet-POS*s, randomly chosen from the synsets we create. Each volunteer was requested to evaluate using a 5-point scale – 5: excellent, 4: good, 3: average, 2: fair and 1: bad. The average score of Wordnet synsets for *arb, asm* and *vie* are 3.82, 3.78 and 3.75, respectively. We notice that the Wordnet synsets generated using the IW approach with all 4 intermediate Wordnets have the highest average score: 4.16/5.00 for *arb* and 4.26/5.00 for *vie*. We are in the process of finding volunteers to evaluate the Wordnet synsets for *ajz* and *dis*.

It is difficult to compare Wordnets because the languages involved in different papers are different, the number and quality of input resources vary and the evaluation methods are not standard. However, for the sake of completeness, we make an attempt at comparing our results with published papers. Although our score is not in terms of percentage, we obtain the average score of 3.78/5.00 (or informally and possibly incorrectly, 75.60% precision) which we believe it is better than 55.30% obtained by (Bond et al., 2008) and 43.20% obtained by (Charoenporn et al., 2008). In addition, the average coverage percentage of all Wordnet synsets we create is 44.85% which is better than 12% in (Charoenporn et al., 2008) and 33276 synsets ($\simeq$ 28.28%) in (Saveski and Trajkovsk, 2010) .

The previous studies need more than one dictionary to translate between a target language and intermediate-helper languages. For example, to create the JWN, (Bond et al., 2008) needs the Japanese-Multilingual dictionary, Japanese-English lexicon and Japanese-English life science dictionary. For *asm*, there are a number of Dict(eng,asm); to the best of our knowledge only two online dictionaries, both between *eng* and *asm*, are available. The IWND approach requires only one input dictionary between a pair of languages. This is a strength of our method.

4 Conclusion and future work

We present approaches to create Wordnet synsets for languages using available Wordnets, a public MT and a single bilingual dictionary. We create Wordnet synsets with good accuracy and high coverage for languages with low resources (*arb* and *vie*), resource-poor (*asm*) and endangered (*ajz* and *dis*). We believe that our work has the potential to construct full Wordnets for languages which do not have many existing resources. We are in the process of creating a Website where all Wordnet synsets we create will be available, along with a user friendly interface to give feedback on individual entries. We will solicit feedback from communities that use these languages as mother-tongue. Our goal is to use this feedback to improve the quality of the Wordnet synsets. Some of Wordnet synsets we created can be downloaded from http://cs.uccs.edu/∼linclab/projects.html.

References

Antoni Oliver and Salvador Climent. 2012. Parallel corpora for Wordnet construction: Machine translation vs. automatic sense tagging. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, volume part II, pages 110-121, New Delhi, India, March.

Benoît Sagot and Darja Fišer. 2008. Building a free French Wordnet from multilingual resources. In *Proceedings of the Ontolex 2008 Workshop*, Marrakech, Morocco, May.

Fellbaum, Christiane. 1998. *Wordnet: An electronic lexical database*. MIT Press, Cambridge, Massachusetts, USA.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362, Sofia, Bulgaria, August.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki and Kiyotaka Uchimoto. 2008. Boot-strapping a Wordnet using multiple existing Wordnets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 1619–1624, Genoa, Italy, May.

Eduard Barbu and Verginica Barbu Mititelu. 2005. Automatic building of Wordnets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, September.

Gunawan and Andy Saputra. 2010. Building synsets for Indonesian Wordnet with monolingual lexical resources. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 297–300, Harbin, China, December.

Hiroyuki Kaji and Mariko Watanabe. 2006. Automatic construction of Japanese Wordnet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1262–1267, Genoa, Italy, May.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki. 2008. Development of Japanese Wordnet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2420–2423, Marrakech, Morocco, May.

Krister Lindén and Laur Carlson. 2010. FinnWordnet - WordNet påfinska via översättning, LexicoNordica. *Nordic Journal of Lexicography*, 17:119–140.

Martin Saveski and Igor Trajkovsk. 2010. Automatic construction of Wordnets by using machine translation and language modeling. In *Proceedings of the 13th Multiconference Information Society*, Ljubljana, Slovenia.

Orhan Bilgin, Özlem Çentinoğlu and Kemal Oflazer. 2004. Building a Wordnet for Turkish. *Romanian Journal of Information Science and Technology*, 7(1-2): 163–172.

Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Thatsanee Charoenporn, Virach Sornlertlamvanich, Chumpol Mokarat and Hitoshi Isahara. 2008. Semi-automatic compilation of Asian Wordnet, In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 1041–1044, Tokyo, Japan.