Expanding wordnets to new languages with multilingual sense disambiguation

Expanding wordnets to new languages with multilingual sense disambiguation

Mihael Arcan **John P. McCrae** **Paul Buitelaar**

Insight Centre for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org

Abstract

Princeton WordNet is one of the most important resources for natural language processing, but is only available for English. While it has been translated using the *expand* approach to many other languages, this is an expensive manual process. Therefore it would be beneficial to have a high-quality automatic translation approach that would support NLP techniques, which rely on WordNet in new languages. The translation of wordnets is fundamentally complex because of the need to translate all senses of a word including low frequency senses, which is very challenging for current machine translation approaches. For this reason we leverage existing translations of WordNet in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora. We evaluate our approach using 10 translated wordnets for European languages. Our experiment shows a significant improvement over translation without any contextual information. Furthermore, we evaluate how the choice of pivot languages affects performance of multilingual word sense disambiguation.

1 Introduction

Princeton WordNet (Fellbaum, 1998) is a manually created resource that has been used in many different tasks and applications across linguistics and natural language processing. WordNet's hierarchical structure makes it a useful tool for many semantic applications and it also plays a vital role in modern deep learning based NLP systems (Rychalska et al., 2016). However, Princeton WordNet is only available for English and huge efforts have been made to extend WordNet with multilingual information in projects, such as EuroWordNet (Vossen, 1998), BalkaNet (Tufiş et al., 2004) and MultiWordNet (Pianta et al., 2002). However, most of the wordnet resources resulting from these efforts have fewer synsets than the Princeton WordNet and there are still many languages for which a wordnet does not exist or is not available to all potential users due to licensing restrictions, impacting applications in information retrieval, word sense disambiguation, sentiment analysis or knowledge management that rely on Princeton WordNet.

Most wordnets in languages other than English have followed an *extend* approach (Vossen, 2005), where the structure of Princeton WordNet is preserved and only the words in each synset are translated and new synsets are added for concepts, which are not lexicalized in English. Since manual multilingual translation and evaluation of wordnets using this approach is a very time consuming and expensive process, we apply statistical machine translation (SMT) to automatically translate WordNet entries. While an SMT system can only return the most frequent translation when given a term by itself, it has been observed that SMT provides strong word sense disambiguation when the word is given in the context of a sentence. As a motivating example, we consider the word *vessel*, which is a member of three synsets in Princeton WordNet, whereby the most frequent translation, e.g., as given by Google Translate, is *Schiff* in German and *nave* in Italian, corresponding to `i60833`[1] 'a craft designed for water transportation'. For the second sense, `i65336` 'a tube in which a body fluid circulates', we assume that we know the

[1] We use the CILI identifiers for synsets (Bond et al., 2016)

German translation for this sense is *Gefäß*. In our approach we look for sentences in a parallel corpus, where the words *vessel* and *Gefäß* both occur and obtain a context such as 'blood vessel' that allows the SMT system to translate this sense correctly. This alone is not sufficient as *Gefäß* is also a translation of `i60834` 'an object used as a container', however in Italian these two senses are distinct (*vaso* and *recipiente* respectively), thus by using as many languages as possible we maximize our chances of finding a well disambiguated context.

In this work, we propose an approach to select the most relevant sentences from a parallel corpus based on the overlap with existing translations of WordNet in as many pivot languages as possible. The goal is to identify sentences that share the same semantic information in respect to the synset of the WordNet entry that we want to translate. This approach will allow us to provide a large multilingual WordNet in more than 20 different European languages, which we call Polylingual WordNet.[2] We present multiple evaluations of our approach and show that in general at least 4 languages should be used to assist in the selection of contexts and that languages closely related to the target language should be used in preference to more distant languages. We evaluated our approach on translating WordNet entries into Italian, Slovene, Spanish and Italian, showing improvements between 5 and more than 10 BLEU points compared to a generic translation approach. This approach has been used to expand wordnets for many European languages as well as generate the first wordnet for Maltese.

2 Related Work

Princeton WordNet inspired many researchers to create similarly structured wordnets for other languages. The EuroWordNet project (Vossen, 1998) linked wordnets in different languages through a socalled Inter-Lingual-Index (ILI) into a single multilingual lexical resource. Via this index, the languages are aligned between each other, which allows to go from a concept in one language to a concept with a similar meaning in any of the other languages. Further multilingual extensions were generated by the BalkaNet project (Tufiş et al., 2004), focusing on the Balkan languages and MultiWordNet (Pianta et al., 2002), aligning Italian concepts to English equivalents.

Due to the large interest in the multilingual extensions of the Princeton WordNet, several initiatives started with the aim to unifying and making these wordnets easily accessible. The KYOTO project (Fellbaum and Vossen, 2012) focused on the development of a language-independent module to which all existing wordnets can be connected, which would allow a better cross-lingual machine processing of lexical information. Recently this has been realized by a new Global WordNet Grid (Vossen et al., 2016) that takes advantage of the Collaborative Inter-Lingual Index (CILI) (Bond et al., 2016). Since most of the current non-English wordnets use the Princeton WordNet as a pivot resource, concepts, which are not in this English lexical resource cannot not be realized or aligned to it. Therefore the authors support the idea of a central platform of concepts, where new concepts may be added even if they are not represented (yet) in the Princeton WordNet or even lexicalized in English (e.g., many languages have distinct gendered role words, such as 'male teacher' and 'female teacher', but these meanings are not distinguished in English).

Previous studies of generating non-English wordnets combined Wiktionary knowledge with existing wordnets to extend them or to create new ones (de Melo and Weikum, 2009). Bond and Paik (2012) describe in their work the creation of the Open Multilingual Wordnet and its extension with other resources (Bond and Foster, 2013). A different approach to expand English WordNet synsets with lexicalizations in other languages was proposed in de Melo and Weikum (2012). The authors do not directly match concepts in the two different language resources, but demonstrate an approach that learns how to determine the best translation for English synsets by taking bilingual dictionaries, structural information of the English WordNet and corpus frequency information into account. With the growing amount of parallel data, Kazakov and Shahid (2009) show an approach to acquire a set of synsets from parallel corpora. The synsets are obtained by comparing aligned words in parallel corpora in several languages. Similarly, the sloWNet for Slovene (Fišer, 2007) and Wolf for French (Sagot and Fišer, 2008) are constructed using a multilingual corpus and word alignment techniques in combination with other existing lexical resources.

[2]The Polylingual WordNet is available at `http://polylingwn.linguistic-lod.org/`

Since all these approaches use word alignment information, they are not able to generate any translation equivalents for multi-word expressions (MWE). In contrast, our approach use an SMT system trained on a large amount of parallel sentences, which allows us to align possible MWEs, such as *commercial loan* or *take a breath*, between source and target language. Furthermore, we engage the idea of identifying relevant contextual information to support an SMT system translating short expressions, which showed better performance compared to approaches without a context. Arcan et al. (2015) built small domain-specific translation models for ontology translation from relevant sentence pairs that were identified in a parallel corpus based on the ontology labels to be translated. With this approach they improve the translation quality over the usage of large generic translation models. Since the generation of translation models can be computational expensive, Arcan et al. (2016) use large generic translation models to translate ontology labels, which were placed into a disambiguated context. With this approach the authors demonstrate translation quality improvement over commercial systems, like *Microsoft Translator*. Different from this approach, which uses the hierarchical structure of the ontology for disambiguation, we engage a large number of different languages to identify the relevant context.

Oliver and Climent (2012) present a method for WordNet construction and enlargement with the help of sense tagged parallel corpora. Since parallel sense tagged data are not always available, they use *Google Translate* to translate a manually sense tagged corpus. In addition they apply automatic sense tagging of a manually translated parallel corpus, whereby they report worse performance compared to the previous approach. We try to overcome this issue by engaging up to ten languages to improve the performance of the automatic sense tagging. Similarly, BabelNet (Navigli and Ponzetto, 2012) aligns the lexicographic knowledge from WordNet to the encyclopaedic knowledge of Wikipedia. This is done by assigning WordNet synsets to Wikipedia entries, and making these relations multilingual through the interlingual links. For languages, which do not have the corresponding Wikipedia entry, the authors use *Google Translate* to translate English sentences containing the synset in the sense annotated corpus. After that, the most frequent translation is included as a variant for the synset for the given language.

The use of parallel corpora has been previously exploited for word sense disambiguation, for example to construct sense-tagged corpora in another language (Ng et al., 2003) or by using translations as a method to discriminate senses (Ide et al., 2002). It has been shown that the combination of these techniques can improve supervised word sense disambiguation (Chan et al., 2007). A similar approach to the one proposed in this paper is that of Tufiş et al. (2004), where they show that using the interlingual index of WordNet with the help of parallel text can improve word sense disambiguation of a monolingual approach and we generalize this result to generate wordnets for new languages.

3 Methodology

Our approach takes the advantage of the increasing amount of parallel corpora in combination with wordnets in languages other than English for sense disambiguation, which will help us to improve automatic translations of English WordNet entries. We assume that we have a multilingual parallel corpus consisting of sentences, x_i^l in a language l, grouped into parallel translations:

$$\mathcal{X} = \{(x_i^{l_0}, \ldots, x_i^{l_T})\}$$

We also assume that we have a collection of wordnets consisting of a set of senses, w_{ij}^l, grouped into synsets, for each language:

$$\mathcal{S} = \{(\{w_{ij}^{l_0}\}, \ldots, \{w_{ij}^{l_T}\})\}$$

We say that a context $x_i^{l_0}$, in language l_0 (in our case this is always English), is *disambiguated in* n *languages* for a word $w_{jk}^{l_0}$ if:

$$\exists w_{jk_1}^{l_1}, \ldots, w_{jk_n}^{l_n} : w_{jk_1}^{l_1} \in x_i^{l_1} \wedge \ldots \wedge w_{jk_n}^{l_n} \in x_i^{l_n}$$

That is, a context is disambiguated in n languages for a word, if for each of its translations we have a context in the parallel corpus that contains one of the known synset translations. Furthermore, we assume

we have an SMT system that can translate any context in l_0 into our target language, l_T, and produces a phrase alignment such that we know which word in the output corresponds to the input word. We used the following methods to choose contexts for the SMT system:

None The SMT system is given only the word $w_{jk}^{l_0}$ as a single sentence as input, thus the most frequent translation is returned.

Random context A random $x_i \in \mathcal{X}$, such that $w_{jk}^{l_0} \in x_i^{l_0}$, is chosen.

Disambiguated context The contexts are ordered by the number of languages that they are disambiguated in, and the context that is disambiguated in the maximal number of languages is chosen. If there are multiple such languages, one context is chosen at random.

m **Disambiguated contexts** The contexts are ordered, as above, and the m top scoring contexts are used, with ties broken at random. Each of these contexts is given to the SMT system and the most frequent translation across these m contexts is used. The previous mode is the same as this when $m = 1$.

t**-best Translations** The SMT system is configured to return the t highest scoring translations, according to its model, and we select the translation as the most frequent translation of the context among this t-best list. In our experiments, we combined this with m disambiguations to give tm candidate translations from which the candidate is chosen.

Target Side Lookup (TSL) We can also utilize the translation of our context into the target language $x_i^{l_T}$ from the parallel corpus, however this cannot be applied directly as we do not know which word(s) in $x_i^{l_T}$ correspond to the input and previous work (Arcan et al., 2014) has shown that automatic inference of this alignment (e.g., with GIZA++) can seriously affect performance. Instead we filter contexts to those that generate a translation candidate, $w_k^{l_T}$, such that $w_k^{l_T} \in x_i^{l_T}$, i.e., the machine translation agrees with the gold-standard translation for this context.

4 Experimental Setting

This section gives an overview on the multilingual resources and the translation toolkit used in our experiment. Furthermore, we give insights into SMT evaluation techniques, considering the translation direction of the English WordNet entries into Italian, Slovene, Spanish and Croatian.

4.1 Wordnets for Sense Disambiguation in Parallel Corpora

Princeton WordNet is a large, publicly available lexical semantic database of English nouns, verbs, adjectives and adverbs, grouped into synsets ($\approx$ 117,000). We engage further wordnets in a variety of languages, provided by the Open Multilingual Wordnet web page.[3] The individual wordnets have been made by many projects and we use ten wordnets in different languages for our experiments, i.e, Croatian (Oliver et al., 2015), Dutch (Postma et al., 2016), Finnish (Lindén and Carlson., 2010), French (Sagot and Fišer, 2008), Italian (Toral et al., 2010), Polish (Maziarz et al., 2012), Portuguese (de Paiva and Rademaker, 2012), Romanian (Tufiş et al., 2008), Slovene (Fišer et al., 2012) and Spanish (Gonzalez-Agirre et al., 2012) WordNet. Table 1 illustrates the size of the wordnets and their coverage compared to the Princeton WordNet (last row).[4]

4.2 Statistical Machine Translation

Our approach is based on phrase-based SMT (Koehn et al., 2003), where we wish to find the best translation of a string, given by a log-linear model combining a set of features. The translation that maximizes the score of the log-linear model is obtained by searching all possible translations candidates.

[3] `http://compling.hss.ntu.edu.sg/omw/`

[4] Core refers to the percentage of synsets covered from the semi-automatically compiled list of 5000 "core" word senses in Princeton WordNet.

Language	Synsets	Words	Senses	Core	Language	Synsets	Words	Senses	Core
Croatian	23,120	29,008	47,900	100%	Polish	33,826	45,387	52,378	54%
Dutch	30,177	43,077	60,259	67%	Portuguese	43,895	54,071	74,012	84%
Finnish	116,763	129,839	189,227	100%	Romanian	56,026	49,987	84,638	94%
French	59,091	55,373	102,671	92%	Slovene	42,583	40,233	70,947	86%
Italian	35,001	41,855	63,133	83%	Spanish	38,512	36,681	57,764	76%

Table 1: Statistics on used wordnets for sense disambiguation on parallel corpora.

Parallel Corpus (language pair)	Source Words	Target Words	Parallel Sentences	Parallel Corpus (language pair)	Source Words	Target Words	Parallel Sentences
English–Croatian[1,2]	165M	133M	16M	English-Polish[2]	361M	296M	34M
English–Dutch[2]	426M	372M	37M	English–Portuguese[2]	391M	377M	33M
English–Finnish[2]	248M	165M	25M	English–Slovene[1,2]	166M	130M	13M
English–French[2]	730M	784M	52M	English–Spanish[1,2]	391M	378M	37M
English–Italian[1,2]	273M	270M	22M	English–Romanian[2]	317M	302M	43M

Table 2: Statistics on parallel data for translation model training and word-sense disambiguation. (parallel resources used for training the translation models[1] and/or word-sense disambiguation[2])

The decoder, which is a search procedure, provides the most probable translation based on a statistical translation model learned from the training data.

For our translation task, we use the statistical translation toolkit Moses (Koehn et al., 2007), where word alignments, necessary for generating translation models, were built with the GIZA++ toolkit (Och and Ney, 2003). The Kenlm toolkit (Heafield, 2011) was used to build a 5-gram language model.

4.3 Parallel Resources for SMT training and Word-Sense-Disambiguation

To ensure a broad lexical and domain coverage of our SMT system we merged the existing parallel corpora for each language pair from the OPUS web page[5] into one parallel data set, i.e., Europarl (Koehn, 2005), DGT - translation memories generated by the *Directorate-General for Translation* (Steinberger et al., 2014), MultiUN corpus (Eisele and Chen, 2010), EMEA, KDE4, OpenOffice (Tiedemann, 2009), OpenSubtitles2012 (Tiedemann, 2012). Similarly, we concatenate parallel corpora for identifying relevant sentences containing WordNet entries, which are then translated into the targeted languages. Table 2 shows the number of parallel sentences used for the ten language pairs.

4.4 Translation Evaluation Metrics

The automatic translation evaluation is based on the correspondence between the SMT output and reference translation (gold standard). For the automatic evaluation we used the BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and chrF (Popović, 2015) metrics. **BLEU** (Bilingual Evaluation Understudy) is calculated for individual translated segments (n-grams) by comparing them with a data set of reference translations.[6] The calculated scores, between 0 and 100 (perfect translation), are averaged over the whole *evaluation data set* to reach an estimate of the translation's overall quality. Considering the short length of the terms in WordNet, while we report scores based on the unigram overlap (BLEU-1), this is in most cases only precision, so in addition we also report other metrics. **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching. In contrast to

[5] `http://opus.lingfil.uu.se/index.php`

[6] Due to the possibility of including multiple references for evaluation within the BLUE metric, we use the set of target words within a synset as our gold standard.

	types		tokens	
	Num.	Perc.	Num.	Perc.
English-Italian	507	6.3	521	4.2
English-Spanish	396	4.9	406	3.3
English-Slovene	633	7.9	656	5.3
English-Croatian	600	7.5	621	5.0

Table 3: Number of Out-Of-Vocabulary words and their percentage between translation models and WordNet senses.

	Number	Percentage
English-Italian	4239	40.35
English-Spanish	4436	42.22
English-Slovene	4523	43.04
English-Croatian	3986	37.94

Table 4: Statistics (actual number and percentage) of identified context for the evaluated WordNet Senses.

BLEU, the metric produces good correlation with human judgement at the sentence or segment level. **chrF3** is a character n-gram metric, which has shown very good correlations with human judgements on the WMT2015 shared metric task (Stanojević et al., 2015), especially when translating from English into morphologically rich(er) languages. As there are multiple translations available for each sense in the target wordnet we use all translations as multiple references for BLEU, for the other two metrics we compare only to the most frequent member of the synset.

The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant with a p-value < 0.05.

5 Evaluation

In this section we present the evaluation of the translated English WordNet words into Italian, Slovene, Spanish and Croatian. We evaluate the quality of translations of the WordNet entries based on the provided contextual information as well as the impact on the number of languages and their effect on word-sense disambiguation.

5.1 Translation Quality Evaluation Based on Contextual Information

Our main evaluation focuses on the importance of identifying relevant contexts for translation into Spanish, Italian, Slovene and Croatian. For a comparable evaluation we translated only senses within synsets, which exist in all four targeted languages. Due to the large parallel corpora used to build the translation models, only a small percentage of the used senses (10,507) could not be translated (Table 3). For this evaluation, we required contexts to be disambiguated by at least five out of nine[7] other languages. For around 40% of these senses we could identify relevant context, which was used to guide the SMT to translate the WordNet senses in the right domain (Table 4).

Table 5 illustrates the contribution of the provided contextual information, which supports the SMT system in translating the WordNet entries into the correct sense. We observed that translating a WordNet entry without any contextual information, which we consider as our baseline, provides better translations than translating them within a random context, as the most frequent translation is more likely to be correct than a random disambiguation. Once we identify one unambiguous sentence with a WordNet entry to be translated, the translation quality significantly improves in terms of the BLEU metric for all four targeted languages. Due to the large amount of parallel resources (between ≈15 and ≈50 Million sentences) we provide further a set of ten disambiguated sentences to the SMT system and select the most frequent translation of the targeted English WordNet entry. We observed, that the usage of most frequent translation helps us to improve the translation quality for 1.1 (for Slovene) and 0.7 (for Croatian) BLEU score points. In our last setting we provide the most frequent translation out of the set of t-best possible translations provided by the SMT system, however this does not seem to increase the quality of translation. Finally, in all settings we applied the target side lookup (TSL) procedure and found that it improves the quality of translation in nearly all settings.

[7]The target language is not used to help for sense disambiguation.

		English→Spanish			English→Slovene		
Context	TSL	BLEU-1	METEOR	chrF	BLEU-1	METEOR	chrF
None (baseline)	/	65.8	33.0	64.0	49.4	21.2	56.3
Random	no	54.4	27.2	61.3	36.9	15.7	52.8
Random	yes	53.0	26.6	59.3	36.4	15.9	52.4
Disambiguated	no	66.2	32.4	65.7	52.9	22.8	57.5
Disambiguated	yes	67.8	33.5	64.6	56.0	24.7	58.1
10 disambiguated	no	65.9	32.2	65.5	54.0	23.5	57.9
10 disambiguated	yes	**70.8**	**35.0**	**66.6**	**57.9**	25.4	59.0
5-best 10 disambiguated	no	66.8	32.7	65.9	55.0	23.5	57.1
5-best 10 disambiguated	yes	68.8	33.8	64.7	57.1	**28.4**	**59.6**
		English→Italian			English→Croatian		
Context	TSL	BLEU-1	METEOR	chrF	BLEU-1	METEOR	chrF
None (baseline)	/	62.5	28.4	59.6	51.1	23.8	60.7
Random	no	46.4	20.6	56.1	40.3	18.4	56.9
Random	yes	49.4	21.3	56.4	39.5	17.9	54.9
Disambiguated	no	61.5	26.6	61.7	55.1	24.7	60.0
Disambiguated	yes	66.1	27.8	61.6	57.8	26.5	60.8
10 disambiguated	no	61.0	26.2	61.3	55.8	25.6	61.1
10 disambiguated	yes	**68.0**	**28.6**	**62.7**	**61.4**	**28.3**	**63.2**
5-best 10 disambiguated	no	63.1	27.2	61.8	56.7	25.2	60.7
5-best 10 disambiguated	yes	67.1	28.2	61.6	58.7	27.1	61.5

Table 5: Evaluation of WordNet translations into Spanish, Slovene, Italian and Croatian with context-aware techniques (TSL = Target Sentence Lookup; number of languages used for disambiguation = 5)

Error Analysis In order to investigate to what extent the automatically generated translations differ from the existing entries in the target wordnets we manually inspected the WordNet translations. We compare results where contextual information was used with the approach where WordNet entries were translated in isolation, hence without context. For Slovene, the contextual information provided a correct translation of the WordNet entry *space* (outer space/location outside the Earth's atmosphere, i81724) as *vesolje*, where the context-less translation approach produced the word *prostor*, in the meaning of place, room or property. Similarly, translating *medicine* (medical science, i38643) without contextual information provided a wrong translation as *zdravilo* (medication, drug, i56119), instead of the Slovene equivalent *medicina*. For Italian, an evident mistake was observed when translating the word *tip* (gratuity, i106560), where the translation of the word in isolation wrongly produced *punta*, meaning "the top or extreme point of something" (i82274). A correct translation in Italian supported by the contextual information was provided as *mancia*. Further, *union*, in the meaning of trade union or brotherhood (i80384), *sindacato* in Italian, was wrongly translated into the most dominant meaning *unione*, with its meaning of combination or cohesion. In Croatian, the word *weed* (i105476) as "any plant that crowds out cultivated plants", was wrongly translated into *trava* (drug street name, i57595), if translated in isolation. The correct translation as *korov* was generated with the disambiguated contextual information. For Spanish, *town* (i82504) was mistranslated into *ciudad* (city or large town), whereby the preferred sense of the translation *pueblo* (small town) was generated by using the contextual information.

5.2 Impact of the Number of Languages for Sense Disambiguation

Even with a very large parallel corpus, as we increase the number of languages, in which we disambiguate the sense, we find that for many senses we cannot find a context that is disambiguated in all languages. Thus, we evaluate the impact of changing the number of languages used to disambiguate an English

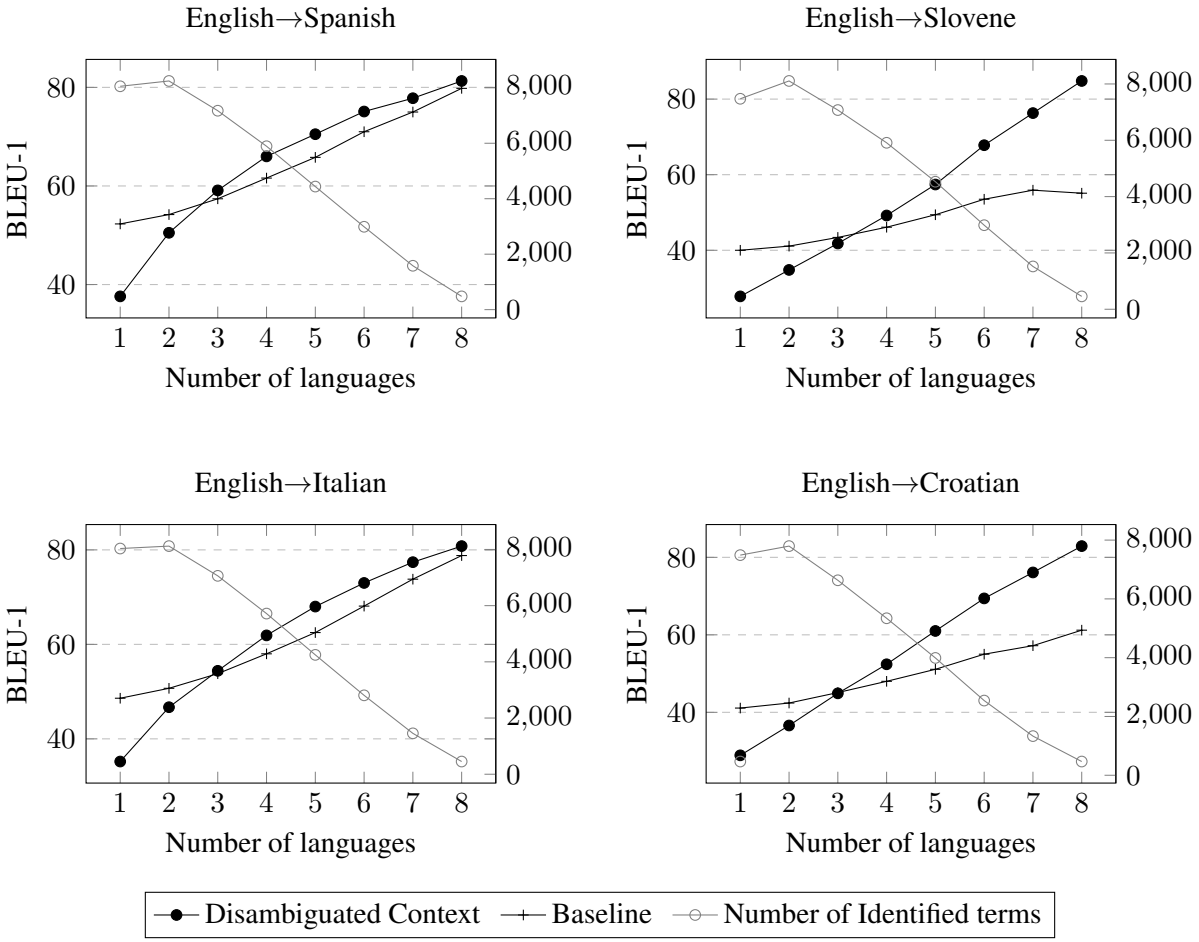

Figure 1: Impact of languages used for disambiguation and translation quality in terms of BLEU.

sentence. For this experiment we report the BLEU scores obtained by the best approach identified in Section 5.1, i.e. *10 disambiguated contexts*. For this evaluation we steadily increase the number of languages that we require a sense to be disambiguated in. We compare these results to the baseline setting, where WordNet entries are translated without any context. As the total number of senses that can be translated decreases, the BLEU score for the baseline does not stay constant and in fact increases, as the senses that our method can disambiguate in many languages are those that are more frequent and less ambiguous. Nevertheless, the disambiguation outperforms the baseline if the context is disambiguated in more than three languages (Figure 1).

For the Romance languages (Italian and Spanish), we outperform the baseline between 3 and 6 BLEU points. The improvement is more evident for the Slavic languages (Slovene and Croatian), where the differences can reach more than 10 BLEU points, if five or more languages are used. For all targeted languages, the observed improvements are statistically significant ($p<0.005$).

5.3 Impact of Language Family for Sense Disambiguation

In addition to the evaluation based on disambiguated contextual information and number of different languages, we were interested in how the similarity of languages affects the disambiguation. Firstly, we focus on the translations of English, a Germanic language, into Slovene, which is a member of the Slavic language family. We considered the cases, where the context is disambiguated in four languages, but looked at two different sets of four languages. Firstly, a group where four languages are of the same family, but different to the source and target language, using four Romance languages: French, Spanish, Romanian and Portuguese. Secondly, we evaluate the sense disambiguation approach using two Romance languages, French and Spanish, and two Slavic languages, Croatian and Polish. As illustrated

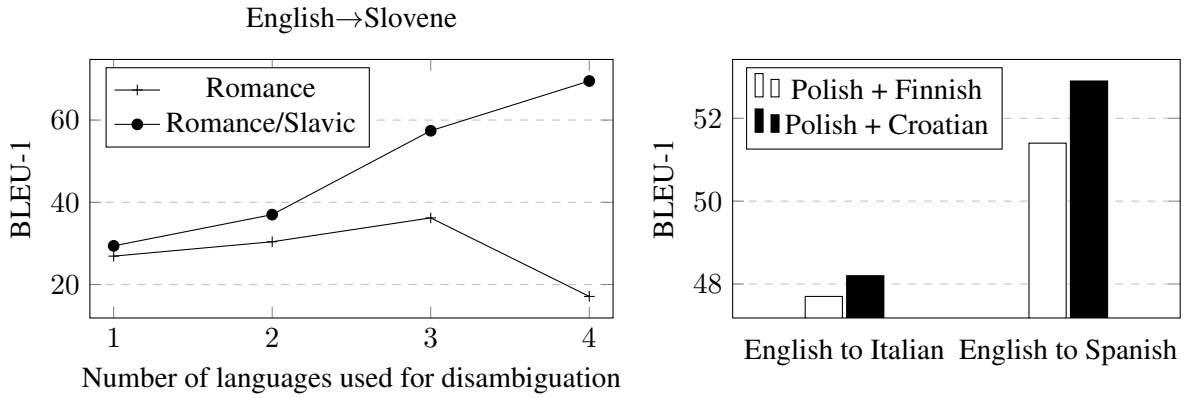

Figure 2: Evaluation on impact of closely-related languages on sense disambiguation for translation quality.

in Figure 2 (left part) the contextual disambiguation approach work significantly better if languages, closely related to the target language – in our case Slovene – are used. In our scenario including the Slavic languages to disambiguate the context yields to better translation quality compared to the usage of only Romance languages.

Secondly, we evaluated our approach if a very distant language is used in the disambiguation, namely Finnish, which is not part of the Indo-European family, the super-family of Romance, Germanic and Slavic languages. We perform disambiguation using Polish and Finnish and compare the results when Finnish is replaced with the Croatian language. The results in Figure 2 (right part) show that Finnish has less disambiguation power than Croatian even though Croatian is similar to Polish. This is because Croatian, even though it is not close to Spanish or Italian is still much closer than Finnish is.

This experiment showed that closely related languages contribute in the disambiguation approach, which yields in our scenario to better translation quality. They also show that using a diverse selection of highly distinct languages does not seem to be advantageous in disambiguating senses.

6 Conclusion and Future Work

We showed an automatic approach to increase the coverage of WordNet into different languages with high-quality translations. By identifying disambiguated context, we demonstrate statistical significant translation improvement for Spanish, Italian, Slovene and Croatian. We demonstrate the importance on closely related languages used for the sense disambiguation approach, which will help us in our ongoing work on generating translations of wordnets beyond the four targeted languages used in this work. This method allows us to release high quality extensions of Princeton WordNet, expanding the coverage for many languages, as well as creating wordnets for languages, where no wordnet has been created or the wordnet is not available to all potential users due to licensing issues.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight) and the European Union supported project MixedEmotions (H2020-644632).

References

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2014. Enhancing statistical machine translation with bilingual terminology in a CAT environment. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, Vancouver, Canada.

Mihael Arcan, Marco Turchi, and Paul Buitelaar. 2015. Knowledge portability with semantic expansion of ontology labels. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, China, July.

Mihael Arcan, Mauro Dragoni, and Paul Buitelaar. 2016. Translating ontologies in real-world settings. In *Proceedings of the 15th International Semantic Web Conference (ISWC-2016)*, Osaka, Japan.

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 (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan. 64–71.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*, Bucharest, Romania.

Yee Seng Chan, Hwee Tou Ng, and Zhi Zhong. 2007. NUS-PT: exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 253–256. Association for Computational Linguistics.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability . In *Proceedings of the Association for Computational Lingustics*.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Gerard de Melo and Gerhard Weikum. 2012. Constructing and utilizing wordnets using statistical methods. *Language Resources and Evaluation*, 46(2):287–311.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Lang. Resour. Eval.*, 46(2):313–326, June.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Darja Fišer. 2007. Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. In *Language and Technology Conference*, pages 359–368. Springer Berlin Heidelberg.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117, Matsue, Japan. The Global WordNet Association.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66. Association for Computational Linguistics.

Dimitar Kazakov and Ahmad R. Shahid. 2009. Unsupervised construction of a multilingual wordnet from parallel corpora. In *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, MCTLLL '09, pages 9–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Antoni Oliver and Salvador Climent. 2012. Parallel corpora for wordnet construction: Machine translation vs. automatic sense tagging. In *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*, pages 110–121.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2015. Automatic expansion of Croatian Wordnet. In *In Proceedings of the 29th CALS international conference "Applied Linguistic Research and Methodology"*, Zadar (Croatia).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP team at SemEval-2016 task 1: Necessity for methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 614–620.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.

Jörg Tiedemann. 2012. Character-based pivot translations for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France, April.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai, India.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian WordNet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452, Szeged.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1312. Association for Computational Linguistics.

Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference (GWC2016)*, Bucharest, Romania.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Piek Vossen. 2005. Building wordnets. `http://www.globalwordnet.org/gwa/BuildingWordnets.ppt`.