WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation

Bharathi Raja Chakravarthi, Mihael Arcan, John P. McCrae
Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
bharathi.raja@insight-centre.org,
mihael.arcan@insight-centre.org, john@mccr.ae

Abstract

In this paper, we translate the glosses in the English WordNet based on the expand approach for improving and generating wordnets with the help of multilingual neural machine translation. Neural Machine Translation (NMT) has recently been applied to many tasks in natural language processing, leading to state-of-the-art performance. However, the performance of NMT often suffers from low resource scenarios where large corpora cannot be obtained. Using training data from closely related language have proven to be invaluable for improving performance. In this paper, we describe how we trained multilingual NMT from closely related language utilizing phonetic transcription for Dravidian languages. We report the evaluation result of the generated wordnets sense in terms of precision. By comparing to the recently proposed approach, we show improvement in terms of precision.

1 Introduction

Wordnets are lexical resource organized as hierarchical structure based on synset and semantic features of the words (Miller, 1995; Fellbaum, 1998). Manually constructing wordnet is a difficult task and it takes years of experts' time. Another way is translating synsets of existing wordnet to the target language, then applying methods to identify exact matches or providing the translated synset to linguists and this has been proven to speed up wordnet creation. The latter approach is known as the *expand* approach. Popular wordnets like EuroWordNet (Vossen, 1997) and IndoWordNet (Bhattacharyya, 2010) were based on the *expand* approach. On the Global WordNet Association website,[1] a comprehensive list of wordnets available for different languages can be found, including IndoWordNet and EuroWordNet.

Due to the lack of parallel corpora, machine translation systems for less-resourced languages are not readily available. We attempt to utilize Multilingual Neural Machine Translation (MNMT) (Ha et al., 2016), where multiple sources and target languages are trained simultaneously without changes to the network architecture. This has been shown to improve the translation quality, however, most of the under-resourced languages use different scripts which limits the application of these multilingual NMT. In order to overcome this, we transliterate the languages on the target side and bring it into a single script to take advantage of multilingual NMT for closely-related languages. Closely-related languages refer to languages that share similar lexical and structural properties due to sharing a common ancestor (Popović et al., 2016). Frequently, languages in contact with other language or closely-related languages like the Dravidian, Indo-Aryan, and Slavic share words from a common root (*cognates*), which are highly semantically and phonologically similar.

In the scope of the wordnet creation for under-resourced languages, combining parallel corpus from closely related languages, phonetic transcription of the corpus and creating multilingual neural machine translation has been shown to improve the results in this paper. The evaluation results ob-

[1] http://globalwordnet.org/

tained from MNMT with transliterated corpus are better than the results of Statistical Machine Translation (SMT) from the recent work (Chakravarthi et al., 2018).

2 Related Work

The Princeton WordNet (Miller, 1995; Fellbaum, 1998) was built from scratch. The taxonomies of the languages, synsets, relations among synset are built first in the merge approach. Popular wordnets like EuroWordNet (Vossen, 1997) and IndoWordNet (Bhattacharyya, 2010) are developed by the expand approach whereby the synsets are built in correspondence with the existing wordnet synsets by translation. For the Tamil language, Rajendran et al. (2002) proposed a design template for the Tamil wordnet.

To evaluate and improve the wordnets for the targeted under-resourced Dravidian languages, Chakravarthi et al. (2018) followed the approach of Arcan et al. (2016), which uses the existing translations of wordnets in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora. They use this contextual information to improve the translation quality of WordNet senses. They showed that their approach can help overcome the drawbacks of simple translations of words without context. Chakravarthi et al. (2018) removed the code-mixing based on the script of the parallel corpus to reduce the noise in translation. The authors used the SMT to create bilingual MT for three Dravidian languages. In our work, we use MNMT system and we transliterate the closely related language corpus into a single script to take advantage of MNMT systems.

Neural Machine Translation achieved rapid development in recent years, however, conventional NMT (Bahdanau et al., 2015) creates a separate machine translation system for each pair of languages. Creating individual machine translation system for many languages is resource consuming, considering there are around 7000 languages in the world. Recent work on NMT, specifically on low-resource (Zoph et al., 2016; Chen et al., 2017) or zero-resource machine translation (Johnson et al., 2017; Firat et al., 2016) uses third languages as pivots and showed that translation quality is significantly improved. Ha et al. (2016) proposed an approach to extend the Bahdanau et al. (2015) architecture to multilingual translation by sharing the entire model. The approach of shared vocabulary across multiple languages resulted in a shared embedding space. Although the results were promising, the result of the experiments was reported in highly resourced languages such as English, German, and French but many under-resourced languages have different syntax and semantic structure to these languages. Chakravarthi et al. (2019) shown that using languages belonging to the same family and phonetic transcription of parallel corpus to a single script improves the MNMT results.

Our approach extends that of Chakravarthi et al. (2019) and Chakravarthi et al. (2018) by utilizing MNMT with a transliterated parallel corpus of closely related languages to create wordnet sense for Dravidian languages. In particular, we downloaded the data, removed code-mixing and phonetically transcribed each corpus to Latin script. Two types of experiments were performed: In the first one, where we just removed code-mixing and compiled the multilingual corpora by concatenating the parallel corpora from three languages. In the second one removed code-mixing, phonetically transcribed the corpora and then compiled the multilingual corpora by concatenating the parallel corpora from three languages. These two experiments are contribution to this work compared to the previous works.

3 Experiment Setup

3.1 Dravidian Languages

For our study, we perform experiments on Tamil (ISO 639-1: ta), Telugu (ISO 639-1: te) and Kannada (ISO 639-1: kn). The targeted languages for this work differ in their orthographies due to historical reasons and whether they adopted the Sanskrit tradition or not (Bhanuprasad and Svenson, 2008). Each of these has been assigned a unique block in Unicode, and thus from an MNMT perspective are completely distinct.

3.2 Multilingual Neural Machine Translation

Johnson et al. (2017) and Ha et al. (2016) extended the architecture of Bahdanau et al. (2015) to use a universal model to handle multiple source and target languages with a special tag in the encoder to determine which target language to translate. The idea is to use the unified vocabulary and training corpus without modification in the architecture to take advantage of the shared embedding. The goal of this approach is to improve the translation quality for individual languages pairs, for which parallel corpus data is scarce by letting the NMT to learn the common semantics across languages and reduce the number of translation systems needed. The sentence of different languages are distinguished through languages codes.

3.3 Data

We used datasets from Chakravarthi et al. (2018) in our experiment. The authors collected three Dravidian languages $\leftrightarrow$ English pairs from OPUS[2] web-page (Tiedemann and Nygaard, 2004). Corpus statistics are shown in Table 1. More descriptions about the three datasets can be found in Chakravarthi et al. (2018). We transliterated this corpus using Indic-trans library[3]. All the sentences are first tokenized with OpenNMT (Klein et al., 2017) tokenizer and then segmented into subword symbols using Byte Pair Encoding (BPE) (Sennrich et al., 2016). We learn the BPE merge operations across all the languages. Following Ha et al. (2016), we indicate the language by prepending two tokens to indicate the desired source and target language. An example of a sentence in English to be translated into Tamil would be:

```
src__en tgt_ta I like ice-cream
```

3.4 Transliteration

As the Indian languages under our study are written in different scripts, they must be converted to some common representation before training the MNMT to take advantage of closely related language resources. A phonetic transcription is an approach where a word in one script is transformed into a different script by maintaining phonetic correspondence. Phonetic transcribing to Latin script and International Phonetic Alphabet (IPA) was studied by (Chakravarthi et al., 2019) and showed that Latin script outperforms IPA for the MNMT Dravidian languages. The improvements in results were shown in terms of the BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and chrF (Popović, 2015) metric. To evaluate the similarity of the corpus the authors used cosine similarity and shown that transcribing to Latin script retain more similarity. We used Indic-trans library by Bhat et al. (2015), which bring all the languages into a single representation by phoneme matching algorithm. The same library can also back-transliterate from English (Latin script) to Indian languages.

[2] http://opus.nlpl.eu/

[3] https://github.com/libindic/indic-trans

3.5 Code-Mixing

Code-mixing is a phenomenon which occurs commonly in most multilingual societies where the speaker or writer alternate between two or more languages in a sentence (Ayeomoni, 2006; Ranjan et al., 2016; Yoder et al., 2017; Parshad et al., 2016). Since most of our corpus came from publicly available parallel corpus are created by voluntary annotators or align automatically. The technical documents translation such as KDE, GNOME, and Ubuntu translations have code-mixing data since some of the technical terms may not be known to voluntary annotators for translation. But the code-mixing from OpenSubtitle are due to bilingual and historical reasons of Indian speakers (Chanda et al., 2016; Parshad et al., 2016). Different combinations of languages may occur while code-mixing for example German-Italian and French-Italian in Switzerland, Hindi-Telugu in state of Telangana, India, Taiwanese-Mandarin Chinese in Taiwan (Chan et al., 2009). Since the Internet era, English become the international language of the younger generation. Hence, English words are frequently embedded in Indians' speech. For our work, only intra-sentential code-mixing was taken into account. In this case, Dravidian languages as the primary language, and English as secondary languages. We removed the English words considering only the English as a foreign word based on the script. Statistics of the removal of code-mixing is shown in Table 2.

3.6 WordNet creation

Using contextual information to improve the translation quality of wordnet senses was shown to improve the results (Arcan et al., 2016). The approach is to select the most relevant sentences from a parallel corpus based on the overlap of existing wordnet translations. For each synset of wordnet entry, multiple sentences were collected that share semantic information. We use this contextual data in English to be translated into Tamil, Telugu, and Kannada using our MNMT system.

4 Results

We present consolidated results in Table 3. Apart from Precision at 1, the Table 3 shows Precision at 2, Precision at 5, Precision at 10. The goal of

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
Number of tokens	7,738,432	6,196,245	258,165	226,264	68,197	71,697
Number of unique words	134,486	459,620	18,455	28,140	7,740	15,683
Average word length	4.2	7.0	3.7	4.8	4.5	6.0
Average sentence length	5.2	7.9	4.6	5.6	5.3	6.8
Number of sentences	449,337		44,588		13,543	

Table 1: Statistics of the parallel corpora used to train the translation systems.

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
tok	0.5% (45,847)	1.1% (72,833)	2.8% (7,303)	4.9% (12,818)	3.5% (2,425)	9.0% (6,463)
sent	0.9% (4,100)		3.1% (1,388)		3.4% (468)	

Table 2: Number of sentences (sent) and number of tokens (tok) removed from the original corpus.

this work is to aid the human annotator in speeding up the process of wordnet creation for under-resourced languages. Precision at different levels is calculated by comparing it with IndoWordNet for the exact match out of the top 10 words from word alignment based on the attention model in MNMT and alignment from SMT. The precision of all the MNMT systems is greater than the baseline.

The perfect match of a word and IndoWordNet entry is considered for Precision at 1. Tamil, Telugu, and Kannada yield better precision at a different level for translation based on both MNMT. For Tamil and Telugu, the translation based on MNMT trained on the native script and MNMT trained on transcribed script did not have much variance. The slight reduction in the result is caused by the transliteration into and back to the original script. In the case of Kannada, which has very less number of parallel sentences to train compared to the other two languages, the MNMT translation trained on transcribed script shows high improvement.

	English→Tamil			
	P@10	P@5	P@2	P@1
B-SMT	0.1200	0.1087	0.0833	0.0651
NC-SMT	0.1252	0.1147	0.0911	0.0725
NC-MNMT	**0.2030**	**0.1559**	0.1228	0.1161
NCT-MNMT	0.1816	0.1538	**0.1351**	**0.1320**
	English→Telugu			
	P@10	P@5	P@2	P@1
B-SMT	0.0471	0.0455	0.0380	0.0278
NC-SMT	0.0467	0.0451	0.0382	0.0274
NC-MNMT	**0.0933**	0.0789	0.0509	0.0400
NCT-MNMT	0.0918	**0.0807**	**0.0599**	**0.0565**
	English→Kannada			
	P@10	P@5	P@2	P@1
B-SMT	0.0093	0.0096	0.0080	0.0055
NC-SMT	0.0110	0.0107	0.0091	0.0067
NC-MNMT	0.0652	0.0472	0.0319	0.0226
NCT-MNMT	**0.0906**	**0.0760**	**0.0535**	**0.0433**

Table 3: Results of Automatic evaluation of translated wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10. B-Baseline original corpus, NC- Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Multilingual Neural Machine Translation

We have several observations. First, the precision presented is below 15 percent and this is because these languages have very minimum parallel corpora. Chakravarthi et al. (2018) used the corpora collected during August 2017 from OPUS which contains mostly translation of religious text, technical document, and subtitles. Analyzing the results by comparing with IndoWordNet is likely to be problematic since it is far from complete and is overly skewed to the classical words for these languages. Second, our method outperforms the baseline from (Chakravarthi et al., 2018) for all the languages, demonstrating the effectiveness of our framework for multilingual NMT. More importantly, transliterating the parallel corpora is more beneficial for the low resource language pair English-Kannada.

Manual Evaluation

In order to re-confirm the validity of the output in practical scenarios, we also performed a human-based evaluation in comparison with IndoWordNet entries. For human evaluation 50 wordnet entries from the wordnet were randomly selected. All these entries were evaluated according to the manual evaluation method performed by Chakravarthi et al. (2018). The classification from the paper is given below. More details about the classification

	B-SMT	NC-SMT	NC-MNMT	NC-MNMT-T
Agrees with IndoWordNet	18%	20%	**28%**	26%
Inflected form	12%	22%	26%	**30%**
Transliteration	4%	4%	2%	2%
Spelling variant	2%	2%	2%	2%
Correct, but not in IndoWordNet	18%	**24%**	22%	**24%**
Incorrect	46%	28%	20%	**16%**

Table 4: Manual evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage. B-Baseline original corpus, NC- Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Multilingual Neural Machine Translation

can be found in Chakravarthi et al. (2018).

- **Agrees with IndoWordNet** Perfect match with IndoWordNet.
- **Inflected form** Some parts of a word such root of a word is found.
- **Transliteration** Transliteration of an English word in Tamil this might be due to unavailability of the translation in the parallel corpus.
- **Spelling Variant** Spelling variant can be caused by wrong or misspelling of the word according to IndoWordNet. Since the corpus contains data from OpenSubtitle this might include dialect variation of the word.
- **Correct, but not in IndoWordNet** Word sense not found in IndoWordNet but found in our translation. We verified we had identified the correct sense by referring to the wordnet gloss.
- **Incorrect** This error class can be caused due to inappropriate term or mistranslated.

Table 4 contains the percentage for outputs of the wordnet translation. As mentioned earlier in Section 3, SMT systems trained on removing code-mixing and without removing are used as baselines for this assessment. The baseline system shows that the cleaned data (removing code-mix) produce better results. Again, as we previously mentioned both our MNMT system trained on cleaned data are better than the baseline system in the manual evaluation as well. From Table 4, we can see that there is a significant improvement over the inflected form MNMT systems trained with the transcribed corpus. Perfect match with IndoWordNet is lower for MNMT trained with transcribed corpus compared to MNMT trained on the original script but still better than the baselines. This might be due to back-transliteration effect. It is clear from the results that this translation can be used as an aid by annotators to create wordnet for under-resourced languages.

5 Conclusion

In this paper, we presented how to take advantage of phonetic transcription and multilingual NMT to improve the wordnet sense translation of under-resourced languages. The proposed approach incorporates code-mixing phenomenon into consideration as well as the phonetic transcription of closely related language to better utilize multilingual NMT. We evaluated the proposed approach on three Dravidian languages and showed that the proposed approach outperforms the baseline by effectively leveraging the information from closely related languages. Moreover, our approach can provide better translations for very low resourced language pair (English-Kannada). In the future, we would like to conduct an experiment by transcribing the languages to one of the Dravidian languages scripts which will be able to represent information more easily than Latin script.

Acknowledgments

This work is supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

Arcan, Mihael, John P. McCrae, and Paul Buitelaar. 2016. Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of The 26th International Conference on Computational Linguistics*.

Ayeomoni, Moses Omoniyi. 2006. Code-switching and code-mixing: Style of language use in childhood in Yoruba speech community. *Nordic Journal of African Studies*, 15(1):90–99.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Bhanuprasad, Kamadev and Mats Svenson. 2008. Errgrams - A Way to Improving ASR for Highly Inflected Dravidian Languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Bhat, Irshad Ahmad, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Bhattacharyya, Pushpak. 2010. IndoWordNet. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Chakravarthi, Bharathi Raja, Mihael Arcan, and John P. McCrae. 2018. Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global WordNet Conference*. The Global WordNet Conference 2018 Committee.

Chakravarthi, Bharathi Raja, Mihael Arcan, and John P. McCrae. 2019. Comparison of Different Orthographies for Machine Translation of Under-resourced Dravidian Languages. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*.

Chan, Joyce Y. C., Houwei Cao, P. C. Ching, and Tan Lee. 2009. Automatic Recognition of Cantonese-English Code-Mixing Speech. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*, September.

Chanda, Arunavha, Dipankar Das, and Chandan Mazumdar. 2016. Columbia-Jadavpur submission for emnlp 2016 code-switching workshop shared task: System description. *EMNLP 2016*, page 112.

Chen, Yun, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935. Association for Computational Linguistics.

Fellbaum, Christiane, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Firat, Orhan, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277. Association for Computational Linguistics.

Ha, Thanh-Le, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the International Workshop on Spoken Language Translation*.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, December.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

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

Papineni, Kishore, 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 of the Association for Computational Linguistics*.

Parshad, Rana D., Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is India speaking? Exploring the "Hinglish" invasion. *Physica A: Statistical Mechanics and its Applications*, 449:375 – 389.

Popović, Maja, Mihael Arcan, and Filip Klubička. 2016. Language Related Issues for Machine Translation between Closely Related South Slavic Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52. The COLING 2016 Organizing Committee.

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

Rajendran, S, S Arulmozi, B Kumara Shanmugam, S Baskaran, and S Thiagarajan. 2002. Tamil WordNet. In *Proceedings of the First International Global WordNet Conference. Mysore*, volume 152, pages 271–274.

Ranjan, Prakash, Bharathi Raja, Ruba Priyadharshini, and Rakesh Chandra Balabantaray. 2016. A comparative study on code-mixed data of Indian social media vs formal text. In *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 608–611. IEEE.

Tiedemann, Jorg and Lars Nygaard. 2004. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus . In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Vossen, Piek. 1997. EuroWordNet: a multilingual database for information retrieval. In *In: Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

Yoder, Michael, Shruti Rijhwani, Carolyn Rosé, and Lori Levin. 2017. Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 73–82. Association for Computational Linguistics.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575. Association for Computational Linguistics.