NUI Galway
OÉ Gaillimh

ARAN - Access to Research at NUI Galway

Provided by the author(s) and NUI Galway in accordance with publisher policies. Please cite the published version when available.

| | |
|---|---|
| Title | Improving wordnets for under-resourced languages using machine translation |
| Author(s) | Chakravarthi, Bharathi Raja; Arcan, Mihael; McCrae, John P. |
| Publication Date | 2018-01-08 |
| Publication Information | Chakravarthi, Bharathi Raja, Arcan, Mihael, & McCrae, John P. (2018). Improving wordnets for under-resourced languages using machine translation. Paper presented at the GWC 2018, The 9th Global WordNet Conference, Nanyang Technological University (NTU), Singapore, 8 – 12 January. |
| Publisher | Global Wordnet Association |
| Link to publisher's version | http://compling.hss.ntu.edu.sg/events/2018-gwc/#proceedings |
| Item record | http://hdl.handle.net/10379/14878 |

Downloaded 2023-03-23T11:55:30Z

Some rights reserved. For more information, please see the item record link above.

# Improving Wordnets for Under-Resourced Languages Using 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

Wordnets are extensively used in natural language processing, but the current approaches for manually building a wordnet from scratch involves large research groups for a long period of time, which are typically not available for under-resourced languages. Even if wordnet-like resources are available for under-resourced languages, they are often not easily accessible, which can alter the results of applications using these resources. Our proposed method presents an *expand* approach for improving and generating wordnets with the help of machine translation. We apply our methods to improve and extend wordnets for the Dravidian languages, i.e., Tamil, Telugu, Kannada, which are severly under-resourced languages. We report evaluation results of the generated wordnet senses in term of precision for these languages. In addition to that, we carried out a manual evaluation of the translations for the Tamil language, where we demonstrate that our approach can aid in improving wordnet resources for under-resourced Dravidian languages.

## 1 Introduction

As computational activities and the Internet creates a wider multilingual and global community, under-resourced languages acquire political as well as economic interest to develop Natural Language Processing (NLP) systems for these languages. In general, creating NLP systems requires an extensive amount of resources and manual effort, however, under-resourced languages lack in both.

Wordnets are lexical resources, which provide a hierarchical structure based on synsets (a set of one or more synonyms) and semantic features of individual words. Wordnets can be constructed by either the *merge* or the *expand* approach (Vossen, 1997). Princeton WordNet (Miller, 1995; Fellbaum, 2010) was manually created within Princeton University covering the vocabulary in English language only. Then, based on the Princeton WordNet, wordnets for several languages were created. As an example, EuroWordNet (Vossen, 1997) is a multilingual lexical database for several European languages, structured in the same way as Princeton's WordNet. The Multiwordnet (Pianta et al., 2002) is strictly aligned with Princeton WordNet and allows to access senses in Italian, Spanish, Portuguese, Hebrew, Romanian and Latin language. Many others have followed for different languages. The IndoWordNet (Bhattacharyya, 2010) was compiled for eighteen out of the twenty-two official languages of India and made available for public use. It is based on the *expand* approach like EuroWordNet, but from the Hindi wordnet, which is then linked to English. On the Global WordNet Association website,[1] a comprehensive list of wordnets available for different languages can be found, including IndoWordNet and EuroWordNet etc.

This paper describes the effort towards generating and improving wordnets for the under-resourced Dravidian languages. Since studies (Federico et al., 2012; Läubli et al., 2013; Green et al., 2013) have shown significant productivity gains when human translators post-edit machine translation output rather than translating text from scratch, we use the available parallel corpora from multiple sources, like OPUS,[2] to create a machine translation system to translate the wordnet senses in the Princeton WordNet into the mentioned under-resourced languages. Translation tools such as Google Translate,[3] or open source SMT systems such as Moses (Koehn et

[1] http://globalwordnet.org/
[2] http://opus.lingfil.uu.se/
[3] http://translate.google.com/

al., 2007) trained on generic data are the most common solutions, but they often result in unsatisfactory translations of domain-specific expressions. Therefore, we follow the idea of Arcan et al. (2016b), where the authors automatically identify relevant sentences in English containing the WordNet senses and translate them within the context, which showed translation quality improvement of the targeted entries. The effectiveness of our approach is evaluated by comparing the generated translations with the IndoWordNet entries, automatically and manually, respectively. This paper reports our first outcomes in improving wordnet for under-resourced Dravidian languages such as Tamil(ISO 639-2: tam), Telugu (ISO 639-2: tel) and Kannada (ISO 639-2: kan).

## 2 Related work

Scannell (2007) describes the start of the creation of a resource for the Irish language using the Web as a resource for NLP approaches. This work started by creating a resource for Irish language using the Web as a resources for NLP. Since 2000, the author and his collaborators developed many resources like monolingual corpora, bilingual corpora and parsers etc, for many under-resourced languages, but they did not cover all languages in the world. A six-level typology was proposed by Alegria et al. (2011) that separated languages into six levels. According to the authors, except for top ten languages in the world all the other languages are under-resourced languages. The third and fourth level languages are the languages which have some resource on the internet. These six level typologies is a relative definition for the under-resourced language, but still can be useful for our study of under-resourced languages.

IndoWordNet covers official Indian languages, from the major three families: Indo-Aryan, Dravidian and Sino-Tibetan languages. In general, Indian languages are rich in morphology and each of the three language families has different morphology structure. It was compiled for eighteen out of the twenty-two official languages and made publicly available.[4] Similarly to EuroWordNet it is based on the *expand* approach, but the central language is Hindi, which is then linked to English. The IndoWordNet entries are updated frequently. For the Tamil language, Rajendran et al. (2002) proposed a design template for the Tamil wordnet. In their further work (Rajendran et al., 2010), they emphasize the need for an independent wordnet for the Dravidian languages, based on EuroWordNet. This is due the observation that the morphology and lexical concepts of these languages are different compared to other Indian languages. The authors have combined the Tamil wordnet and wordnets in other Dravidian languages to form the IndoWordNet.

Mohanty et al. (2017) built SentiWordNet for the Odia language, which is one of the official languages of India. Being an under-resourced language, Odia lacks proper machine translation system to translate the vocabulary of the available resource from English into Odia. The authors have created SentiWordNet for Odia using resources of other Indian languages and the IndoWordNet. Although the IndoWordNet structure does not map directly to the SentiWordNet, instead synsets are matched. The authors used these for translation from source lexicon to target lexicon. Aliabadi et al. (2014) have created a wordnet for the Kurdish language, one of the under-resourced languages in western Iranian language family. They have created Kurdish translation for the "core" wordnet synsets (Vossen, 1997), which is a set of 5,000 essential concepts. They used a dictionary to translate its literals (words), adopted an indirect evaluation alternative in which they look at the effectiveness of using KurdNet for rewriting Information Retrieval queries. Similarly, the work by Horváth et al. (2016) focuses on the semi-automatic construction of wordnet for the Mansi language, which is spoken by Mansi people in Russia, an endangered under-resourced languages with a low number of native speakers. The authors have used the Hungarian wordnet as a starting point. With the help of a Hungarian-Mansi dictionary, which was used to create possible translations between the languages, the Mansi wordnet was continuously expanded.

Previous works did lots of manual effort to create wordnet-like resources, which was funded by public research for a long period of time. However, IndoWordNet is not complete and biased towards Hindi, because the authors created a Hindi-Tamil bilingual dictionary, rather than a wordnet. As explained in Rajendran et al. (2010), the morphology and lexical concepts of Dravidian languages are different from Hindi, which illustrates that the IndoWordNet may not be the most suitable resource to represent the wordnet for the targeted Dravidian languages.

[4] http://www.cfilt.iitb.ac.in/indowordnet/index.jsp

To evaluate and improve the wordnets for the targeted Dravidian languages, we follow the approach of Arcan et al. (2016b), 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. We use this contextual information to improve the translation quality of WordNet senses. We show that our approach can help overcome drawbacks of simple translations of words without context.

## 3 Background

Our specific aim of this work is to generate and improve wordnets for under-resourced languages. For our task we chose the *expand* approach and automatically translated the Princeton WordNet entries within a disambiguate context to obtain entries for the Dravidian languages.

### 3.1 Dravidian languages

Dravidian languages, a family of languages spoken primarily in the Southern part of India and also spread over South Asia. The Dravidian languages are divided into four groups: South, South-Central, Central, and North groups. Dravidian morphology is agglutinating and exclusively suffixal. Words are built from small elements called morphemes. Two broad classes of morphemes are stems and affixes. Words are made up of morphemes concatenated based on the grammar of language. Tamil language is also a free word-order language. Due to the nature of morphology, the noun phrase and verb phrase may appear in any permutation and still able to produce same sense of the sentence (Steever, 1987).

The four major literary Dravidian languages are Tamil, Telugu, Malayalam, and Kannada. Tamil, Malayalam, and Kannada fall under the South Dravidian subgroup, whereby Telugu belongs to the South Central Dravidian subgroup (Vikram and Urs, 2007). All the four languages have official status in Government of India and use their own unique script. Outside India, Tamil also has official status in Sri Lanka and Singapore. Tamil script is descended from the Southern Brahmi script and has 12 vowels, 18 consonants and one aytam (special sound). The Telugu script is also descendant of the Southern Brahmi script. It has 16 vowels and 36 consonants, which are more in number than those of Tamil alphabets. The Kannada and Telugu scripts are most similar and often considered as a regional variant. The Kannada script is used to write other under-resourced languages like Tulu, Konkani and Sankethi. In the Kannada language, the derivation of words is either by combining two distinct words or by affixes. Different to Tamil, Kannada and Telugu inherits some of the affixes from Sanskrit.

### 3.2 Machine Translation

Statistical Machine Translation (SMT) systems assume that we have a set of example translations($S^{(k)}$, $T^{(k)}$) for $k = 1 \ldots n$, where $S^{(k)}$ is the $k^{th}$ source sentence, $T^{(k)}$ is the $k^{th}$ target sentence which is the translation of $S^{(k)}$ in the corpus. SMT systems try to maximize the conditional probability $p(t|s)$ of target sentence $t$ given a source sentence $s$ by maximizing separately a language model $p(t)$ and the inverse translation model $p(s|t)$. A language model assigns a probability $p(t)$ for any sentence t and translation model assigns a conditional probability $p(s|t)$ to source / target pair of sentence. By Bayes rule

$$p(t|s) \propto p(t)p(s|t) \quad (1)$$

This decomposition into a translation and a language model improves the fluency of generated texts by making full use of available corpora. The language model is not only meant to ensure a fluent output, but also supports difficult decisions about word order and word translation (Koehn, 2010). We used the Moses (Koehn et al., 2007) toolkit that provides end-to-end support for the creation and evaluation of machine translation system based on BLEU (Papineni et al., 2002) score. There are two major criteria for automatic SMT evaluation: completeness and correctness, which are considered by BLEU, an automatic evaluation technique, which is a geometric mean of n-gram precision. BLEU score is language independent, fast, and shows good correlation with human evaluation campaigns. Therefore we plan to use this metric to evaluate our work.

### 3.3 Available Corpora for Machine Translation

This section describes the data collection and the pre-processing process steps. The English-Tamil parallel corpus, which we used to train our SMT system is collected from various sources and combined into a single parallel corpus. We used the EnTam corpus (Ramasamy et al., 2012), which was pre-processed from raw Web data to become a sentence-aligned corpus. The parallel corpora

| | 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.

contains text from the news domain,[5] sentences from the Tamil cinema articles[6] and the Bible.[7] For the news corpus, the authors downloaded web pages that have matching file names in both English and Tamil. For the cinema corpus, all the English articles had a link to the corresponding Tamil translation. The collection of the Bible corpus followed a similar pattern. We also took the English-Tamil parallel corpora for six Indian languages created with the help of Mechanical Turk for Wikipedia documents (Post et al., 2012). Since the data was created by non-expert translators hired over the Mechanical Turk, it is of mixed quality. From the OPUS website, we have collected the Gnome, KDE, Ubuntu and movie subtitles (Tiedemann, 2012). We furthermore manually aligned Tamil text Tirukkural,[8] and combined all the parallel corpora into a single corpus. We first tokenized sentences in English and Tamil and then true-cased only the English side of the parallel corpus, since the Tamil language does not have a casing. Finally, we cleaned up the data by eliminating the sentences whose length is above 80 words.

To obtain the parallel corpora for Telugu and Kannada, we used the corpora available on the OPUS website. The same pre-processing procedure was followed for Telugu and Kannada language, since both languages are close to the Tamil language. The Table 1 shows the statistics of the parallel corpora for the three language pairs. From this table we can see that the English-Tamil parallel corpus is much larger than for the other language pairs. On the other hand, the number of sentences for English-Kannada is very small. Once we have obtained the parallel corpus, we created the SMT systems for the English-Tamil, English-Telugu, and English-Kannada language pairs.

We define the following set of data:

- Development set: Randomly selected 2000 sentences from the parallel corpus as development set is used to measure the system performance of the phrase-based translation model.
- Test set: A blind set of 1000 sentence randomly chosen from parallel corpus that is used to the test the system. There is no overlap between these set of data.
- Training set: A larger size parallel corpus that is used to train the phrase-based translation model. It is remaining corpus after development and test are extracted.

In this work, we focus on three languages from Dravidian family namely, Tamil, Telugu, and Kannada. This is mainly due to available parallel corpora and we believe that this method can be extended for other under-resourced languages without much effort.

### 3.4 Resource Scarceness

There are few resources, which can be used to automatically create a wordnet for under-resourced languages. One way to cross the language barrier is with the help of machine translation. As with any machine learning methods, SMT tends to improve translation quality when using a large amount of training data. That is, if the training method sees a specific word or phrase multiple times during training, it is more likely to learn a correct translation. SMT suffers due to the scarcity of parallel corpora, Dravidian word order and the morphological complexity attached to the language. For the Dravidian languages when translating from or to English the translation models suffer because of syntactic differences while the morphological differences contribute to data sparsity. In contrast, small corpora used for training lead to incomplete word coverage, which may cause the out-of-vocabulary (OOV) issues.

Besides the resource scarceness, another issue observed with the corpus for Dravidian languages was code-switching contents in the data. Code-switching is an act of alternating between elements of two or more languages, which is prevalent in

[5] http://www.wsws.org/
[6] http://www.cinesouth.com/
[7] http://biblephone.intercer.net/
[8] http://www.projectmadurai.org/

| | Original | Non-Code mixing |
|---|---|---|
| English→Tamil | 20.29 | 20.61 |
| English→Telugu | 28.81 | 28.25 |
| English→Kannada | 14.64 | 14.45 |

Table 2: Automatic translation evaluation of the of 1000 randomly selected sentences in terms of the BLEU metric.

multilingual countries (Barman et al., 2014). With English being the most used language in the digital world, people tend to mix English words with their native languages. That might be the case in other languages as well.

## 4 Methodology

The principle approaches for constructing wordnets are the *merge* approach or the *expand* approach. In the *merge* approach, the synsets and relations are built independently and then aligned with WordNet. The drawbacks of the *merge* approach are that it is time-consuming and requires a lot of manual effort to build. On the contrary in the *expand* model, wordnet can be created automatically by translating synsets using different strategies, whereby the synsets are built in correspondence with the existing wordnet synsets. We followed the *expand* approach and created a machine translation systems to translate the sentences, which contained the WordNet senses in English to the target language

### 4.1 Training Machine Translation parameters

In the following section, we takes as a baseline a parallel text, that has been aligned at the sentence level. To obtain the translations, we use Moses SMT toolkit with of baseline setup with 5-gram language model created using the training data by KenLM (Heafield, 2011). The baseline SMT system was built for three language pairs, English-Tamil, English-Telugu, and English-Kannada. The test set mentioned in Section 3.3 was used to evaluate our system. From Table 1 and Table 2 we can see that size of the parallel corpus has an impact on the BLEU score for test set which is evaluation criteria for the translation model.

### 4.2 Context Identification

Since manual translation of wordnets using the extend approach is a very time consuming and expensive process, we apply SMT to automatically translate WordNet entries into the targeted Dravidian languages. While an domain-unadapted SMT system can only return the most frequent translation when given a term by itself, it has been observed that translation quality of single word expressions improves when the word is given in an disambiguated context of a sentence (Arcan et al., 2016a; Arcan et al., 2016b). Therefore existing translations of WordNet senses in other languages than English were used to select the most relevant sentences for wordnet senses from a large set of generic parallel corpora. 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. To ensure a broad lexical and domain coverage of English sentences, existing parallel corpora for various language pairs were merged 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, wordnets in a variety of languages, provided by the Open Multilingual Wordnet web page,[9] were used.

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`[10] '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 German translation for this sense is *Gefäß* and we look in our approach 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.

### 4.3 Code-mixing

Code-switching and code-mixing is a phenomenon found among bilingual communities all

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

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

| | 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 3: Number of sentences (sent) and number of tokens (tok) removed from the original corpus.

Source sentence: "இப்போது, நான் அதை loving."
Transliteration: :Ippōtu, nāṉ atai loving
Target sentence: "Right now, I'm loving it."

Source sentence: "முன்னிருப்பு GNOME பொருள்"
Transliteration: :Muṉṉiruppu GNOME poruḷ
Target sentence: "Default GNOME Theme"

Figure 1: Examples of Code-mixing in Tamil-English parallel corpus. In the first example the verb *loving* is code-mixed in Tamil. In Second Example the noun *GNOME* is code-mixed.

over the world (Ayeomoni, 2006; Yoder et al., 2017). Code-mixing is mixing of words, phrases, and sentence from two or more languages with in the same sentence or between sentences. In many bilingual or multilingual communities like India, Hong Kong, Malaysia or Singapore, language interaction often happens in which two or more languages are mixed. Furthermore, it increasingly occurs in monolingual cultures due to globalization. In many contexts and domains, English is mixed with native languages within their utterance than in the past due to Internet boom. Due to the history and popularity of the English language, on the Internet Indian languages are more frequently mixed with English than other native languages (Chanda et al., 2016).

A major part of our corpora comes from movie subtitles and technical documents, which makes it even more prone to code-mixing of English in the Dravidian languages. In our corpus, movie speeches are transcribed to text and they differ from that in other written genres: the vocabulary is informal, non-linguistics sounds like *ah*, and mixing of scripts in case of English and native languages (Tiedemann, 2008). Two example of code-switching are demonstrated in Figure 1.The parallel corpus is initially segregated into English script and native script. All of the annotations are done using an automatic process. All words from a language other than the native script of our experiment are taken out on both sides of corpus if it occurs in native language side of the parallel corpus. The sentences are removed from both sides if the target language side does not contain native script words in it. Table 3 show the percentage of code-mixed text removed from original corpus. The goal of this approach is to investigate whether code-mixing criteria and corresponding training are directly related to the improvement of the translation quality measured with automatic evaluation and manual evaluation. We assumed that code-mixed text can be found by different scripts and did not evaluate the code-mixing written in the native script or Latin script to write the native language as was done by (Das and Gambäck, 2013)

## 5 Evaluation

The most reliable method to evaluate the wordnet is a manual evaluation, but a manual evaluation of whole the WordNet is time consuming and very expensive. Therefore, we did the automatic evaluation of the our translations and measured the precision. In order to determine the correctness of our work, we have furthermore randomly taken 50 WordNet entries for manual evaluation on these entries.

### 5.1 Automatic Evaluation

In this paper, we have compared our result to the IndoWordNet. Once the translation step the of disambiguated context, containing the target entries, was finished, we use the word alignment information to extract the translation of the WordNet entry. Since several disambiguated sentences per WordNet entry were used, we took the translations for each context and then combined the results to count the most frequent one. The top 10-words entries were compared to the IndoWordNet for the exact match.

We took precision at 10, precision at 5, precision at 2, and precision at 1. We did this comparison for the all the three languages, i.e. Tamil, Telugu, and Kannada. As an additional experiment, we removed the code-mixing part of the corpus and created an new translation system, which was used again to translate the same WordNet entries. The table 4 shows the result of the automatic evaluation of the translation of the entries into the Targeted Dravidian languages. The table shows the precision at the different level of

| | English→Tamil | | | |
|---|---|---|---|---|
| | P@10 | P@5 | P@2 | P@1 |
| original corpus | 0.120 | 0.109 | 0.083 | 0.065 |
| non-code mixed | 0.125 | 0.115 | 0.091 | 0.073 |
| | English→Telugu | | | |
| | P@10 | P@5 | P@2 | P@1 |
| original corpus | 0.047 | 0.046 | 0.038 | 0.028 |
| non-code mixed | 0.047 | 0.045 | 0.038 | 0.027 |
| | English→Kannada | | | |
| | P@10 | P@5 | P@2 | P@1 |
| original corpus | 0.009 | 0.010 | 0.008 | 0.005 |
| non-code mixed | 0.011 | 0.011 | 0.009 | 0.007 |

Table 4: Results of Automatic evaluation of wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10.

the translations, based on the translation model, generation from the original corpus and non-code mixed corpus. Non-code mixed often outperforms the baseline in terms of precision, whereby the difference is less visible in Telugu language. This is likely due to the short sentences in the Telugu corpus. These differences in the precision are significant in the manual evaluation of Tamil tests with 50 samples. The wide difference between manual and automatics evaluation can be explained in part by different forms. Table 4 shows an example of how our system differs from the baseline SMT system and how it benefits the wordnet translation. This is a clear evidence that an SMT without code-mixing described above achieves an improvement over the baseline without using any additional training data. However, it has been shown in Arcan et al. (2016b) that better performance on WordNet translation can be achieved, if the corpora contained a sufficient amount of parallel sentences. Their translation evaluation based on the BLEU metric on unigrams (similar to precision at 1, P@1), showed a range between 0.55 and 0.70 BLEU points, for the well resourced languages, like Slovene, Spanish, Croatian and Italian. Restricting the task to a small data set tends to hurt the translation performance, but it can useful to aid in the creation or improvement of new resources for the under-resourced languages.

### 5.2 Manual Evaluation

In order to able to evaluate our method in contrast to stand-alone approaches, we manually evaluated our method in comparison with IndoWordNet entries. To select the sample for manual evaluation,

| | Original | Non-Code mixing |
|---|---|---|
| Agrees with IWN | 18% | 20% |
| Inflected Form | 12% | 22% |
| Transliteration | 4% | 4% |
| Spelling variant | 2% | 2% |
| Correct, but not in IWN | 18% | 24% |
| Incorrect | 46% | 28% |

Table 5: Manual Evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage.

we proceeded as follows: we randomly extracted a sample of 50 wordnet entries from the WordNet. First, each of these 50 wordnet entries were compared to the IndoWordNet for the exact match. Subsequently, regardless of this decision, each of the 50 wordnet entries were evaluated and classified according to its quality. The classification is the following:

- **Agrees with IndoWordNet** Exact match found in IndoWordNet.
- **Inflected form** The root of a word is found with a different inflection, which can make the translation correct but imprecise.
- **Transliteration** The word is transliterated, which can be caused by the unavailability of the translation form in the parallel corpus, since some words are used in transliteration because of foreign words.
- **Spelling Variant** Since our data in day to day language of Tamil and IndoWordNet is skewed towards classical sense of language. Our method produces the Spelling Variant which can be caused by wrong or misspelling of the word according to IndoWordNet.
- **Correct, but not in IndoWordNet** IndoWordNet is large and it covers eighteen languages, but it lacks some wordnet entries for the Dravidian languages. 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 mistranslation.

The examples in the Figure 2 list the Tamil translation wordnet in our experiment. Neither the word nor its translation has appeared in the training corpus therefore, the SMT system cannot translate the word and chooses to produce the word in English. On the other side, these examples may produce some insights into the word.

| ILI code | Gloss | IWN | Meaning | Translation | Meaning | Comments |
|---|---|---|---|---|---|---|
| 14647235-n | any of several compounds containing chlorine and nitrogen; used as an antiseptic in wounds | நைட்ரஜன் | nitrogen | நைதரசன் | nitrogen | Spelling variant |
| 01026095-v | give the name or identifying characteristics of; refer to by name or some other identifying characteristic | பெயரிடு | name, identity | பெயர் | name | Inflected form, different part-of-speech |
| 00461782-n | a game in which balls are rolled at an object or group of objects with the aim of knocking them over or moving them | பந்து | ball | பௌலிங் | bowling | Correct translation, sense missing in IWN |
| 04751305-n | noticeable heterogeneity | பல்வேறு | diverseness, diversity | பல்வேறு | diverseness, diversity | Agrees with IWN |
| 01546111-v | be standing; be upright | தூக்கு | to lift | நிற்க | to stand | correct translation, sense missing in IWN |

Figure 2: Examples of the manual evaluation of Tamil wordnet entries in comparison to the IndoWordNet (IWN).

We should note that this evaluation was carried out for both, original, uncleaned, corpus as well as cleaned corpus (non-code mixing). We observed that the cleaned data produce better results compared to the original data which have many code-mixing entries. From the table 5, we can see that there is a significant improvement over the inflected form and correct but not found in IndoWordNet categories. This shows that our method can help to improve the wordnet entries for under-resourced languages.

## 6 Discussion

While our automatic evaluation results are a little disappointing, and this is perhaps unsurprising in the context of under-resourced languages as there is very little a data availability for these language, our manual evaluation shows that this is far from reality. Evaluating using a resource such as IndoWordNet is always likely to be problematic as the resource is far from complete and does not claim to cover all words in the Dravidian languages studied in this paper. Moreover, IndoWordNet is overly skewed to the the classical words of these languages, but the majority our parallel corpus is day to day conversation texts. Despite the low precision in determining the exact match to the IndoWordNet, our technique yields 48% for precision at 10 in manual evaluation, although the automatic evaluation considering precision at 10 gave only 12%. Our method relays on IndoWordNet for evaluation but IndoWordNet is biased over one particular language, which is Hindi. The resulting wordnet entries, though noisy, is suitable for aiding wordnet creation for under-resourced languages.

The handling of code-mixing in this paper appears to improve the quality of the proposed translation, outperforming the baseline results of wordnet entries once code-mixed was removed from data. Thus we believe that the method presented here still applicable to resource creation of under-resourced languages.

## 7 Conclusion

In this paper we showed the challenges in building wordnet for under-resourced languages and presented that our method can aid the creation or improvement of wordnets for under-resourced languages. We experimented with available data to created SMT systems for three Dravidian languages and used those as a baseline. To improve the results we removed the code-mixed terms from the corpus. Our results indicated that the proposed removing of code-mixed text from the corpus results in gains for the wordnet entries with limited data.

## 8 Acknowledgements

This work was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

## References

Iñaki Alegria, Xabier Artola, Arantza Diaz De Ilarraza, and Kepa Sarasola. 2011. Strategies to develop language technologies for less-resourced languages based on the case of basque.

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building KurdNet, the Kurdish WordNet. In *Proceedings of the 7th Global WordNet Conference (GWC'14)*, pages 1–6.

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

Mihael Arcan, John P McCrae, and Paul Buitelaar. 2016b. Expanding wordnets to new languages with multilingual sense disambiguation. In *International Conference on Computational Linguistics (COLING-2016)*, Osaka, Japan.

Moses Omoniyi Ayeomoni. 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.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 13–23.

Pushpak Bhattacharyya. 2010. Indowordnet. In Nicoletta Calzolari (Conference Chair), 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).

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

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

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. the last language identification frontier? *TAL*, 54(3):41–64.

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.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Csilla Horváth, Ágoston Nagy, Norbert Szilágyi, and Veronika Vincze. 2016. Where bears have the eyes of currant: Towards a mansi wordnet. In *Proceedings of the Eighth Global WordNet Conference*, pages 130–134.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

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

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. *Machine Translation Summit XIV*, page 83.

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

Gaurav Mohanty, Abishek Kannan, and Radhika Mamidi. 2017. Building a sentiwordnet for odia. *EMNLP 2017*, page 143.

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*, pages 311–318. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Int'l Conference on Global WordNet*.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics.

S Rajendran, 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.

S Rajendran, G Shivapratap, V Dhanlakshmi, and KP Soman. 2010. Building a wordnet for Dravidian languages. In *Proceedings of the Global WordNet Conference (GWC 10)*. Citeseer.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for English-Tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Sanford B Steever. 1987. Tamil and the dravidian languages. *The world's major languages*, pages 725–746.

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. 2008. Synchronizing translated movie subtitles. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Jorg 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. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

T. N. Vikram and Shalini R. Urs, 2007. *Development of Prototype Morphological Analyzer for he South Indian Language of Kannada*, pages 109–116. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Michael Miller Yoder, Shruti Rijhwani, Carolyn Penstein Rosé, and Lori Levin. 2017. Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages. *ACL 2017*, page 73.