Machine Translation for Under-resourced Languages

Bharathi Raja Chakravarthi, Mihael Arcan, John P. McCrae
Insight Centre for Data Analytics, National University of Ireland, Galway
name.surname@insight-centre.org

Keywords: Computer Science and Information Technology

Abstract

Phrase-based statistical machine translation performance suffers when models are trained on a small amount of parallel data and/or on morphologically rich languages. Under-resourced languages do not have a large amount of parallel data and some of them are morphologically rich, which makes machine translation even more difficult. In this paper, a research proposal is presented that aims to generate an acceptable quality translation for languages, which suffer from lack of resources.

1. Introduction

The state-of-the-art implementations of open-source toolkits, like Moses [1] improved the developments in SMT by making it possible to build a system trivially that produces a decent translation quality. However, translation quality is poor when translating from English to a morphologically rich language with few parallel corpus resources. Using phrase-based statistical machine translation (SMT) model to a morphologically rich language such as Czech, Finnish, Irish, Tamil or Turkish involves properly ordering the target words and generating it with the proper morphology.

Deep learning is a powerful technique that has achieved high performance on difficult learning tasks and so it has become a popular machine learning tool in different domains such as image and speech recognition. Deep learning based Neural Machine Translation (NMT) has proven to give better results [2] [3] than SMT. Deep neural networks can be applied to machine translation either indirectly or directly, where the indirect application maintains the existing SMT and attempts to improve the SMT submodels such as word alignment, translation rule selection, reordering and structure prediction using language modeling and joint translation prediction. The direct application of deep neural network is to develop a NMT, which builds a single neural network that can be jointly tuned for both encoder-decoder to maximize the translation performance.

2. Research objectives

The goal of this research is to use available data to reach acceptable translation quality for under-resourced languages. The main benefit of this research would be to get acceptable translations for other under-resourced languages. In this process, the following research questions from subsections will be handled:

2.1 Parallel data

To build an SMT system, a large amount of parallel sentences, as well as a monolingual sentences are needed to improve the translation quality. However, available parallel corpora for under-resourced languages are very limited and for some language pair their might be none. Questions:
1. What size of parallel corpus is required to get the acceptable translation?
2. Is it reliable to use closely related languages to aid translation?

2.2 Out of vocabulary (OOV)

Training of machine translation is limited to a fixed vocabulary seen in training, but the translation task is an open vocabulary problem. With a limited amount of data for under-resource languages, the SMT system faces a small amount of vocabulary, which makes the probability of OOV words will be very high.
Questions:
1. How to handle morphologically complex words?
2. Is it acceptable to copy unknown words into the target text?

2.3 Code-Switching

Code-switching is the act of alternating between elements of two or more languages, which is prevalent in many multilingual countries. With English being the most used language in digital world, people tend to mix English with their native languages [4].
Questions:
1. How do we handle the text if it is phonetically written using Roman script, even though the language uses non-Roman script?
2. To what extent can code-switching be tolerated in parallel or monolingual corpus?

3. Methodology

We choose the Tamil language for our preliminary work. Tamil, which is part of the Dravidian language family, is one of the official languages of India, Sri Lanka, Singapore and considered as minority language in Malaysia. It is verb final, relatively free-word order and morphologically rich language. Tamil allows subject and object drop, but requires subject-verb agreement [5]. Usually, Tamil is written in a phonetic, non-Latin script and is considered a under-resourced language [6].

We have collected an English-Tamil parallel corpus from various sources and combined this data. We created a 400,000 parallel corpus [7][8]-[9], which covers texts from Tirukkual, Bible, Quran, cinema, news, movie subtitles, Gnome, KDE, Ubuntu (Table 1).

Dataset	Sentences	En tokens	Ta Tokens
Train	392211	6708938	5006568
Test	1000	18023	13277
Dev	2000	32238	25035

Table 1: Characteristic of Parallel Corpus
(En- English, Ta- Tamil)

We used a system, called Open-Source Neural Machine Translation (OpenNMT) [10] and the MOSES toolkit [1]. Five-gram language model was trained on the target side of the parallel data. The system is tuned with MERT [11]. We tokenize and true case the data with the scripts provided in Moses. Both systems output are evaluated by BLEU (Bilingual Evaluation Under Study) [12].

4. Result and Discussion

With the initial tests, we get results which favor the NMT over SMT, as seen in the Table 2. After working with our preliminary experiment on Tamil language we conclude that NMT give better results compared to phrase based SMT systems.

	Moses	OpenNMT
Ta-En	19.26	**21.05**
En-Ta	21.00	**38.91**

Table 2: BLEU Scores

So far, we have concentrated on gathering the available parallel corpora for under-resourced languages and setting up the baseline machine translation system. The available corpora for Tamil were slightly larger than for other Dravidian languages on OPUS web page [9]. We plan to collect more available parallel corpora for the Dravidian languages and create translation systems in the near future. Since the available data contains many code-switching contents on the Tamil side of the parallel corpus, we suspect that this might be the case also for other Dravidian languages. That might be the reason that we are getting large difference in OpenNMT BLEU score between English-Tamil and Tamil-English. Example of the code-switching sentence from our Tamil parallel corpus.

[1] “முன்னிருப்பு GNOME பொருள்”
“Default GNOME Theme”

[2] “இப்போது, நான் அதை loving.”
“Right now, I'm loving it.”

In the first example, we have the word “GNOME” (Named Entity), which is written in Roman letters on the Tamil side of the parallel corpus. In the second example, we see that the word “loving” is written in Roman letters. This code-switching impacts the results produced by the machine translation systems, causing it to generate both "காதலிக்கிற" or “loving” in different contexts.

Given that our method was producing acceptable quality translation for Tamil language, we plan in the future to extend our research to other languages. However, to improve the performance of the present system, we plan to study the code-switching phenomena more throughly by applying different approaches, like transliterating unknown words and reducing the amount of code-switching data in parallel corpora based on the code mixed index [4].

8. References

[1] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions 2007 Jun 25 (pp. 177-180). Association for Computational Linguistics.

[2] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014 Jun 3.

[3] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In Advances in neural information processing systems 2014 (pp. 3104-3112).

[4] Barman U, Das A, Wagner J, Foster J. Code mixing: A challenge for language identification in the language of social media. EMNLP 2014. 2014 Oct 25;13.

[5] Rajendran S, Arulmozi S, Shanmugam BK, Baskaran S, Thiagarajan S. Tamil WordNet. In Proceedings of the First International Global WordNet Conference. Mysore 2002 (Vol. 152, pp. 271-274).

[6] Pushpananda R, Weerasinghe R, Niranjan M. Sinhala-Tamil Machine Translation: Towards better Translation Quality. In Australasian Language Technology Association Workshop 2014 2014 (Vol. 129, p. 129).

[7] ZdenekŽabokrtský LO. Morphological processing for English-Tamil statistical machine translation. In 24th International Conference on Computational Linguistics 2012 (p. 113).

[8] Post M, Callison-Burch C, Osborne M. Constructing parallel corpora for six Indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation 2012 Jun 7 (pp. 401-409). Association for Computational Linguistics.

[9] Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. In LREC 2012 May 23 (Vol. 2012, pp. 2214-2218).

[10] Klein G, Kim Y, Deng Y, Senellart J, Rush AM. OpenNMT: Open-Source Toolkit for Neural Machine Translation. arXiv preprint arXiv:1701.02810. 2017 Jan 10.

[11] Bertoldi N, Haddow B, Fouet JB. Improved minimum error rate training in Moses. The Prague Bulletin of Mathematical Linguistics. 2009 Jan 1;91:7-16.

[12] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics 2002 Jul 6 (pp. 311-318). Association for Computational Linguistics.