Multiple En-De Approach For Parallel Data Filtering And Mining

Anitha M [1], Kamalesh V N [2]

[1] *PAL training and development Hassan , India*

[2] *Faculty of computer science and engineering, Sreenidhi institute of science and technology, Hyderabad , Telangana India*

***Abstract*—** Neural Machine Translation (NMT) has obtained state-of-the art performance for many language pairs, with only training with parallel data . Target- side multilingual data plays an vital role in increasing fluency for phrase-based statistical machine translation. We study the use of multilingual data for NMT. In this paper, we learn multilingual text embedding and collective approach which can be applied to multiple language pairs .The approach accredit Multilingual NMT using a single model without any increase in parameters, and it is significantly simpler .We learn a joint multilingual sentence embedding and use the distance between sentences in different languages to filter noisy parallel data and mine parallel data in large news collections.

Keywords*—Natural language processing, parallel sentences, parallel sentence extraction , neural machine translation.*

I. INTRODUCTION

It is today's common practice to use distributed representations of words, often called word embeddings, in almost all NLP applications. It has been shown that syntactic and semantic relations can be captured in this embedding space, as in instance [25]. To process sequences of words, ie. sentences or small paragraphs, these word embeddings need to be "combined" into a representation of the whole sequence.Parallel data, also called bitexts, is an vital resource to train neural machine translation systems (NMT). It is usually assumed that the quality of the automatic translations increases with the amount of available training data. However, it was observed that NMT systems are more sensitive to noise than SMT systems, e.g. [6]. Well known Sources of parallel data are Wikipedia, international news and journals . In addition, there are many texts on the Internet which are potential mutual translations, but which need to be identified and aligned. Typical examples are Wikipedia or news collections which report on the same facts in different languages. These collections are usually called comparable corpora.

In this paper we study an unified approach to filter noisy bitexts and to mine bitexts in huge monolingual texts. The main idea is to first learn a joint multilingual sentence embedding. Then, a threshold on the distance between two sentences in this joint embedding space can be used to filter bitexts (distance between source and target sentences), or to mine for additional bitexts (pairwise distances between all source and target sentences). No additional features or classifiers are needed.

II. RELATED WORK

Interlingual translation is a classic method in machine translation. Despite its distinguished history, most practical applications of machine translation have focused on individual language pairs because it was simply too difficult to build a single system that translates reliably from and to several languages. There have been other approaches similar to ours in spirit, but used for very different purposes.

The problem of how to select parts of bitexts has been addressed before, but mainly from the aspect of domain adaptation[25]. It was successfully used in many phrase-based MT systems, but it was reported to be less successful for NMT [23]. Domain adaptation of neural networks via continued training has been shown to be effective for neural language models and in work for neural translation models[17].

For instance, [27](first embed sentences into two separate spaces. Then, a classifier is learned on labeled data to decide whether sentences are parallel or not. Our approach clearly outperforms this technique on the BUCC corpus. [7] use averaged multilingual word embed- dings to calculate a joint embedding of all sentences. However, distances between all sentences are only used to extract a set of potential mutual translation. The decision is based on a different system. In[11]NMT systems for Zh ↔ En are learned using a joint encoder. A sentence representation is obtained as the mean of the last encoder states. Noisy bitexts are filtered based on the distance. In all these works, embeddings are learned for two languages only, while we learn one joint embedding for up to nine languages.

III. NMT TRAINING WITH MONOLINGUAL TRAINING DATA AND BACK TRANSLATION

We see two simple methods to use monolingual training data during training of NMT systems[19], with no changes to

the network architecture. Training examples are provided with dummy source context and the training was successful to some rate .More results were gained through back-translation of monolingual target data. The monolingual target data was converted into the source language, and this synthetic data was treated as an input training data . This monolingual data and the in domain back-t ranslated was again used for training . With the analysis it is inferred that a reduction of overfitting, domain adaptation effects, and improved fluency as factors for the results of using monolingual data for training.

Since in this approach does not require any rectification in the neural network architecture to integrate monolingual training data , this can be easily applied to other NMT systems. Effectiveness of the approach not only changes with the quality of the MT system used for back-translation, but also relies on the amount of monolingual and parallel data , and the scale of overfitting of the baseline model.

IV. COSINE DISTANCE IN A JOINT MULTILINGUAL SENTENCE EMBEDDING SPACE

A. Multilingual sentence embeddings

Complete sentences in different languages are embedded into single joint space, with the aim that the distance in that space contemplate their semantic difference irrespective of the language. There are several works on learning multilingual sentence embeddings which could be used for integrate different languages into one single space([13];[26]).

The idea is to make use of multiple sequential encoders and decoders and the next step is to train them with N-way aligned corpora . Instead of making use of single en- coder for each language , a shared encoder which handles all the input languages. Joint encoders and decoders are already being used in NMT [13] in which a special token is input. In this approach the joint encoder is not provided with input token and has no information at all on the encoded language, or what will be done with the sentence representation. An approach which similar to this was proposed in [10].

The architecture is trained on many languagesof the Europarl corpus with about 2M sentences each. The word embeddings are of size 384 and the hidden layer of the BLSTM is 512-dimensional. The joint encoder is a 3-layer BLSTM. The 1024 dimensional sentence embedding is obtained by max-pooling over the BLSTM outputs. Dropout is set to 0.1. These settings are similar to those reported in [17], with the difference that slight improvement by using a deeper network for the joint en- coder is observed. Once the system is learned, all the BLSTM decoders are discarded and we only use the multilingual BLSTM encoder to embed the sentences into the joint space.

B. Experimental studies:

All the experiments are performed with the freely available Sequence-to Sequence Py- Torch toolkit from Facebook AI Research,called fairseq-py. A convolutional model is implemented which achieves very competitive output[23]. We use this system to show the enhancements obtained by filtering the standard training data and by combining additional mined data.The data used is freely available so that it can be used to train different NMT architectures.

Lang-Pair	Train			Test	
	En	other	aligned	En	other
En-de	400k	414k	9580	397k	414k
En-fr	370k	272k	9086	373k	277k
En-ru	558k	461k	14435	566k	457k
En-zh	89k	95k	1899	90k	92k

Table 1:BUCC evaluation to mine bitexts.

In this work, focus is mainly on translating from English into German using the WMT'14 data. This task was selected for two reasons:

- only a restricted amount of parallel training data is available (4.5M sentences). 2.4M are crawled and aligned sentences (Common Crawl corpus) and 2.1M are high quality human translations.

as studied in other works, we use news test -2014 as test set. We use mteval-v14.pl on untokenized hypothesis to calculate case-sensitive BLEU score in order to follow the standard WMT evaluation setting. In some papers, BLEU is calculated with multi-bleu.perl on tokenized hypothesis. The fairseq-py system is trained with default parameters, but a slightly different pre and post-processing scheme. In particular, we lowercase all data and use a 40k BPE vocabulary [19].All the outcomes are for one single system only. Table 2gives baseline results using the provided data as it is. We differentiate outcomes when training on human labeled data only, i.e News Commentary and with all WMT'14 training data, i.e. human and Common Crawl (total of 4.5M sentences) [23]report a tokenized BLEU score of 25.16 on a slightly different version of news test -2014 as defined in.

- it is the standard to evaluate NMT systems and most of the comparable results are availablle, example[19];[3].Please note that the aim of this paper is not to set a new state-of-the-art in NMT on this data set, but to show relative improvement with respect to a competitive baseline.

Corpus	Human only(Eparl+NC)	All WMT 14 (Eparl+NC+CC)
Number of sents	2.1 M	4.5M
BLEAU	21.87	24.75

Table 2:Baseline results on WMT 14 en-de

C. Filtering Common Crawl

Filtering corpus, first all the sentences are embedded into the single joint space and find the cosine distance between the source English and the German translation provided. Then as a function of the threshold on this distance extract subset with different size.

All	Commas	<50 words	LID
2399k	2144k	2071k	1935k

Table 3: Pre-processing of the Common Crawl corpus before distance-based filtering.

After initial experiments, it was found that some additional steps before calculating the distances (see Table 3) are required and erase sentences with more than 3 or more commas. Those are indeed often enumerations of names, cities, etc. While such sentences maybe useful to train NMT systems. Multilingual distance is not reliable to distinguish list of named entities(nouns).sentences should be limited with less than 50 words; 3) language identification (LID) on source and target sentences are performed. These steps remove 19% of the overall data. Almost 6% of the data seems to have the wrong source or target language.

D. Filter back-translation of monolingual data using multilingual distance

The back-translation of monolingual target data into the source language to produce synthetic parallel text has been previously explored for phrase- based SMT [2]. While our approach is technically similar, synthetic parallel data fulfils novel roles in NMT.

The above approach be used to filter the back-translated monolingual data [19]. The monolingual data translated to source language can be filtered using joint space filtering method as explained in the above experimental section .

V. CONCLUSION

In this article we study the interlingual nature of the context vectors generated by a multilingual neural machine translation system and study their power in the assessment of mono- and cross-language similarity. We have seen that a simple cosine distance in a joint multilingual sentence embedding space can be used to filter noisy parallel data and to mine for bitexts in large news collections. There are many directions to extend this research, in particular to scale-up to larger corpora. The multilingual sentence distance could be also used in MT confidence estimation, or to filter back-translations of monolingual data [19].

REFERENCES

1. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations (ICLR).
2. Nicola Bertoldi and Marcello Federico. 2009. Do- main adaptation for statistical machine translation with monolingual resources. In Proceedings of the Fourth Workshop on Statistical Machine Translation
3. Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Sys- tems for Spoken Language Domains. In Proceed- ings of the International Workshop on Spoken Lan- guage Translation 2015, Da Nang, Vietnam.
4. Luc´ıa Santamar´ıa and Amittai Axelrod. 2017. Data selection with cluster-based language difference mod- els and cynical selection. In IWSLT, pages 137–145.
5. Andoni Azpeitia, Thierry Etchegoyhen, and Eva Mart´ınez Garcia. 2018. Extracting paral- lel sentences from comparable corpora with STACC variants. In BUCC.
6. Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and artificial noise both break neural machine trans- lation.
7. Houda Bouamor and Hassan Sajjad. 2018. H2@bucc18: Parallel sentence extraction from comparable coprora using multlingual sentence embeddings. In BUCC.
8. Sarath Chandar, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2013. Multi- lingual deep learning. In NIPS DL wshop.
9. Junyoung Chunga, Kyunghyun Choa, and Yoshua Bengio. 2016. A character-level decoder without ex- plicit segmentation for neural machine translation.
10. Cristina Espan˜a-Bonet, A da´m Csaba Varga, Alberto Barro´n-Ceden˜o, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual em- beddings and their use in parallel sentence identifi- cation. IEEE Journal of Selected Topics in Signal Processing, pages 1340–1348.
11. Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Feder- mann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on auto- matic chinese to english translation.
12. Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed seman- tics. In ACL, pages 58–68.
13. Melvin Johnson et al. 2016. Google's multilin- gual neural machine translation system: Enabling zero-shot translation. In
14. Chongman Leong, Derek F. Wong, and Lidia S. Chao. 2018. Um-paligner: Neural network-based parallel sentence identification model. I
15. Aditua Mogadala and Achim Rettinger. 2016. Bilin- gual word embeddings from parallel and non- parallel corpora for cross-language classification. In NAACL, pages 692–702.
16. Hieu Pham, Minh-Thang Luong, and Christopher D. Manning. 2015. Learning distributed representa- tions for multilingual text sequences. In Workshop on Vector Space Modeling for NLP.

17. Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neu- ral machine translation. In ACL workshop on Rep- resentation Learning for NLP.
18. Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In ACL, pages 86–96.
19. Yonghui Wu et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.
20. Jeff Johnson, Matthijs Douze, and Herve´ Je´gou. 2017. Billion-scale similarity search with gpus.
21. Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploit- ing non-parallel
22. Gehring et al., 2017; Ashish Vaswani et al., 2017);
23. Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural ma- chine translation. In EMNLP, pages 1400–1410.
24 . Luc´ıa Santamar´ıa and Amittai Axelrod. 2017. Data selection with cluster-based language difference mod- els and cynical selection. In *IWSLT*, pages 137–14
25. Andrea Frome, Grep S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, marc'Aurelio Ranzato, and Thomas Mikolov. 2013. DeViSa:E a deep visual-semantic embedding model. In *NIPS*.
26. Hieu Pham, Minh-Thang Luong, and Christopher D. Manning. 2015. Learning distributed representa- tions for multilingual text sequences. In *Workshop on Vector Space Modeling for NLP*.
27. Francis Gre´goire and Philippe Langlais. 2017. Bucc 2017 shared task: a first attempt toward a deep learn- ing framework for identifying parallel sentences in comparable corpora. In *BUCC*, pages 46–50.