Character-based and Subword-based Neural Machine Translation

Abstract

In this report we address a research question at first: Could we address the out-of-vocabulary or rare word problem and improve the result of translation? Then we use subword-based and character-based neural machine translation design a experiment to explore the research question and introduce the method. At last we discuss and draw conclusion according to the result and give a expecting of furthur work.

1 Address Research Question

In previous work, we implement a baseline whose basic model architecture is an architecture of the encoder-decoder with the attention mechanism, using a bidirectional LSTM. And this neural machine translation model views the sentences of input and output as sequences of words. And in this report, we will explore its application to a new language — Inuit. There is an result of translation using baseline model shown below:

- sentence: 45656
  Src : uqaujjuijitaqaurniarmat isumajaaqtunut ikajuqtuijjutinik iliijaqsuijuvinirnut ammalu uqaujjuijitigut qaujimaniar&utik qanuittut piusituqarijaujutigut inungnut aturijaullaringnirijanginnik ammalu qaujimaniujunik qanuq uqallagiaksaq iliijaaqaqtunut ammalu iliijaqsuangusimajunut ammalu tautuviniujunut
  Ref : there will be counselling provided to the abusers and the counsellor will know about traditional values and know how to talk to the abuser and the victims and the witnesses
  Hyp : in are _UNK _UNK _UNK _UNK _UNK _UNK _UNK _UNK _UNK and _UNK and _UNK and and and about _EOS
- sentence: 45580
  Src : ikajuqsurniq nunaliralaanguniqsanit
  Ref : support for smaller communities
  Hyp : staff housing _UNK _EOS

According to the translation result of the baseline, we could see that there are a lot of _UNK which represents unknown word in the first long sentence and one _UNK in short sentence. Honestly, we get no actual meaning according to the hypotheis from the example above. Although there is just two example (one long and one short) here, actually we found many hypotheises in the result of baseline translation have the _UNK words which seriously affect the quality of translation. Also I find that every Inuit word is very long and I think if this will influence the result of translation. Meanwhile the value of BLEU of baseline model is 7.98. We could find the fact that the result of this kind of neural machine translation is not good enough and compare the references and hypotheises, many hypotheises (whatever long sentence or short sentence) make no sense. Also there are many unknown word in hypotheises.

According to morphology and syntax of Inuit language, I found some interesting properties which may cause the problem of rare word and out-of-vocabulary. "The Inuit language has a very rich morphological system, in which a succession of different morphemes are added to root words (like verb endings in European languages) to indicate things that, in languages like English, would require several words to express. All Inuit language words begin with a root morpheme to which other morphemes are suffixed. The language has hundreds of distinct suffixes. The language has hundreds of distinct suffixes."(Dorais 2010) As a result of this kind of language system, the words of Inuit are very long and mostly unique. The percentage of 92 of all Inuit language's words appear only once according to Canadian corpus 'the Nunavut Hansard', but only a small percentage of words appear only once in most English corpora of similar size. This means the Inuit language has many 'unique' words that means there are many rare words, even words that is out of vocabulary, which is not similar to English and causes the unknown word problem. There is also another language phenomenon of Inuit that most words are very long. These two language phenomenon will affect the quality of translation and increase the difficulty of translation, so we want to solve these problems by exploring the subword-based and character-based neural machine translation.

As for model architecture, I think the model architecture of baseline does not have enough ability to handle the problems above. So that I attempt a new kind of neural machine translation architecture to improve the result of translation.

Hence, in following part of this report I will investigate following research question:
Could we address the out-of-vocabulary or rare word problem and improve the result of translation?

2 Methology

In order to address the out-of-vocabulary or rare word problem and improve the quality of translation. We decide to using subword-based neural machine translation and character-based model with convolutional encoder. And I will compare the result of baseline and these two model by three kinds of measures. The following is some shortcomings of word-based translation that is why we use other two methods.

1. In any language, there is no perfect word segmentation algorithm. The perfect word segmentation algorithm should be able to divide any sentence into lexemes and morphemes. The problem is that the dictionary is often filled with a lot of words that share a lexeme but have different morphology, such as run, runs, ran, running. All of them may exist in the dictionary, each word corresponds to a word vector, but they are clearly shared the same lexeme - run. 2. Using word-based model has unknown word problems and rare word problems, and these problems refer to some of the dictionary in the training set occurs too few times, which result in the training cannot get a good word vector and make them limited in translating languages with rich morphology. Hence, as a result of above, we need to find none word-based model to translate. Then according to the papers(Lee 2016; Sennrich 2015), using subword-based and character-based translation can solve the above problems and avoiding many morphological variables appear in the dictionary.

2.1 Subword-based Model

Subword-based is a simple and effective approach, making the NMT model capable of open-vocabulary translation by encoding rare and unknown words as sequences of subword units according to paper (Sennrich et al 2015). Translation models require mechanisms go below the word level, especially for the languages with productive word formation processes, because there is not always a 1-to-1 correspondence between source and target words. And the subword-cased models may solve the rare words or unknown words problems.

The neural machine translation architecture of this translation system is similarly implemented as baseline, which is consisted of 3-layers encoder-decoder with the attention mechanism and use a bidirectional LSTM. We adapt the Byte Pair Encoding (Gage, 1994) algorithm from merging frequent pairs of bytes to merging characters or character sequences. Firstly, we initialize the symbol vocabulary with the character vocabulary, and represent each word as a sequence of characters with a special end-forward symbol '·'. We iteratively count all symbol pairs and use a new 'AB' to replace the most frequent pair ('A', 'B'), and each merging produces a new symbol to represent a character n-gram. We do not consider pairs that cross word boundaries. The final symbol vocabulary size is equal to the size of the initial vocabulary and the number of merge operations. Finally, we get the subword sequence of input and use it in the above NMT system.

2.2 Character-based Model with Convolutional Encoder

In this kind of character-based NMT model, there is no need to do preprocessing to segment word to "character". We use the basic architecture and the same decoder of the baseline model. We just implement the encoder that uses convolution and pooling layers and highway network layers, which is different from the baseline. For encoder, first, here is an embedding layer which map the source sentence to a sequence of character. Then we implement a variant 3-layer one-dimensional convolution. Each of them uses along consecutive character embedings and Each of them use different kernel size so that could extract different character-level of relatively complete features. Next we apply padding with different stride. Now in order to extract informative character patterns of different lengths we contact the three output and apply ReLU and then max pooling. The output from the convolutional layer is first split into segments of widths, and max-pooling over time is applied to each segment with no overlap. This procedure selects the most salient features to give a segment embedding. A sequence of segment embeddings from the max pooling layer is fed into a highway network that is shown to significantly improve the quality of a character-level language model when used with convolutional layers.

Table 1: Model Setting

No	Model	Conv.setting	Maxpooling Stride	Highway	Encoder	Decoder	Epoch Num
1	Baseline	No	No	No	3 layers 200 LSTM	3 layers 400 LSTM	40
2	Subword-based	No	No	No	3 layers 200 LSTM	3 layers 400 LSTM	40
3	Chracter-based	Contact 3 Conv.dim = (1,1) Kernel = 2,3,5 Stride = 1 Pedding = 1,2,4	3	2 layers	3 layers 200 LSTM	3 layers 400 LSTM	40

2.3 Measures

Clearly explain what you intend to measure and how it relates your hypotheis.
I intend to measure if the method address the rare words (out-of-vocabulary word) or long words problem and if the mothod could improve translation by value of BLEU, CHRF and human evaluation. And we pick model of specific epoch which has the best value of BLEU in 40 epoch.
1. BLEU
BLEU is an algorithm for evaluating the quality of text, which quality is considered to be the correspondence between a machine's output and a human's output. As for calculating the BLEU, scores are calculated by comparing individual translated segments – generally sentences - with a set of good quality reference translations, and then averaged over the whole corpus. But method of BLEU has some limitations that intelligibility or grammatical correctness are not taken into account, and performs badly when used to evaluate the quality of individual sentences. There is an inherent, systemic problem with any metric based on comparing with one or a few reference translations, because in real life, sentences can be translated in many different ways, sometimes with no overlap. Therefore, we need more methods of measuring and analyzing the results comprehensively.

2. CHRF
CHRF is character n-gram's F-score for automatic evaluation of machine translation output. We believe that this score has a potential as a stand-alone metric, because it is similarly to the F-score based on morpheme and takes into account some morpho-syntactic phenomena (Popovic, 2015). In additional, it is different from the related metrics, which is simple and does not require any additional tools and/or knowledge sources. CHRF is absolutely not only language independent, but also tokenization independent. And CHFR shows good correlations with human judgments both on the system as well as on the segment-level. So the value of CHRF is a convincing measuring way.

3. Human Evaluation
It well known that automatic evaluation metrics such as BLEU encourage reference-like translations and do not fully capture true translation quality(Lee et al., 2015). Therefore, we need carry out a human evaluation from (Graham et al., 2016) where we have human assessors rate both adequacy and fluency of each system translation. Adequacy is the degree to which assessors agree that the system translation expresses the meaning of the reference translation. Fluency is evaluated using system translation alone without any reference translation.

2.4 Design Experiments

We design 3 model to do experiment and the setting is shown in Table 1. The implement of subword-based model is similar as the baseline model. The differences is subword-based model added a data preprocessing. In this processing, we use the BPE algorithm which we introduced above for word segmentation. And also we implement a complex CNN-NMT for character-based model the setting we could find from Table 1 and the method of artitecture is introduced above.

And we pick the final trained model at specific epoch steps which has the best value of BLEU until epoch is 40.

3 Result and Analysis

Table 2 is the result of 3 different model including value of the BLEU and relevant CHRF3. We pick the best value of BLEU and CHRF3 fromeAnd following is some example of result of translation.

According to the Table 2 we could get several conclusions: 1. What ever refer to the value of CHRF3 and BLEU, subword-based model performs the best among these three model. 2. Baseline model performs the worst among these three model according to the value of BLEU and CHRF3 3. subword-based performs better

Table 2: Results of Models

No	Model	BLEU	CHRF3
1	Baseline	7.99	22.25
2	Subword-based	9.69	32.07
3	Character-based	8.23	27.66

than character-based model according to the value of BLEU and CHRF3. According to paper(Lee 2016), the character-based model should perform better than subword-based, but here the fact is not true. I think the reason is that we did not use 8-layers of conv. We just use 3 layers. So may be the learning of featrue is not enough.

According to the result of translation below and result of baseline translation above: 1. We observe that the subword-based and character-based models are tied with respect to both adequacy and fluency. And We think the subword-based model's result yields significantly better fluency. The reason of why character is worse than subword is the same as what I discuss above. 2. We find that whatever subword-based and character-based model could address the rare word and out-of-word problem significantly. There is no _UNK word. And after I checked 100 more hypothesis, there still no unknown word. So I think these two method is pretty better for addressing rare word problem. 3. But I find a phenomenon that there are a lot of repeat words in hypothesis which has been still shown on baseline model's result. Hence I think we could explore the reason of appearance of repeating word and attempt to solve it in the future work. 4. We could find that some long sentence influence the translation. Some of results show an inaccurate translation when using subword-based model. The character-level model performs better on these long, concatenative words with ambiguous segmentation.

- sentence: 45656
  Src : uqauj ju ij itaq aur niar mat isumaj a aqtunut ikajuqtuijjut inik ili ij aq su ij uv inir nut ammalu uqauj ju ij itigut qaujim aniar utik qanuittut piusit uqar ijauj utigut inungnut atur ij aullar ing nirij anginnik ammalu qaujiman iujunik qanuq uqall agi aksaq ili ija aq aqtunut ammalu ili ij aq su angusimaj unut ammalu taut uv in iujunut
  Ref : there will be counselling provided to the abusers and the counsellor will know about traditional values and know how to talk to the abuser and the victims and the witnesses
  Hyp1 : there is a ways and and home and one one and and and and and and and and and and
  Hyp2 : That his a wave and and and and one one and and and and and and and and and and

- sentence: 45580
  Src : ikajuqs urniq nunaliralaanguniq sanit
  Ref : support for smaller communities
  Hyp1 : support for smaller review _EOS
  Hyp2 : support for smaller review _EOS

4 Conclusion and Further Work

Through the experiments, we conclude from our results that there is currently a trade-off between generalization to unseen words. we observe better results with larger subword units of the BPE segmentation. We hope that our test set will help in developing and assessing architectures that aim to overcome this trade-off and perform best in respect to both morphology and syntax. With real-world text containing typos and spelling mistakes, the quality of word-based translation would severely drop, as every non-canonical form of a word cannot be represented. On the other hand, a character-level model has a much better chance recovering the original word or sentence. Indeed, our character-based model is robust against a few spelling mistakes. In future work, we could explore the problem of repeating word which I described above and could improve the performance better.

5 Reference

Dorais, L.J., 2010. Language of the Inuit: Syntax, semantics, and society in the Arctic (Vol. 58). McGill-Queen's Press-MQUP.
Ling, W., Trancoso, I., Dyer, C. and Black, A.W., 2015. Character-based neural machine translation. arXiv preprint arXiv:1511.04586.
Lee, J., Cho, K. and Hofmann, T., 2016. Fully Character-Level Neural Machine Translation without Explicit Segmentation. arXiv preprint arXiv:1610.03017.
Sennrich, R., 2016. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. arXiv preprint arXiv:1612.04629.
Chung, J., Cho, K. and Bengio, Y., 2016. A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147.
Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O. and Zaremba, W., 2014. Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206.
Sennrich, R., Haddow, B. and Birch, A., 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.
Gehring, J., Auli, M., Grangier, D. and Dauphin, Y.N., 2016. A Convolutional Encoder Model for Neural Machine Translation. arXiv preprint arXiv:1611.02344.
Popovic, M., 2015, September. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15) (pp. 392-395).
Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

Appendix