Ensemble of Translators with Automatic Selection of the Best Translation – the submission of FOKUS to the WMT 18 biomedical translation task –

Cristian Grozea
Fraunhofer FOKUS
cristian.grozea@fokus.fraunhofer.de

Abstract

This paper describes the system of Fraunhofer FOKUS for the WMT 2018 biomedical translation task. Our approach, described here, was to automatically select the most promising translation from a set of candidates produced with NMT (Transformer) models. We selected the highest fidelity translation of each sentence by using a dictionary, stemming and a set of heuristics. Our method is simple, can use any machine translators, and requires no further training in addition to that already employed to build the NMT models. The downside is that the score did not increase over the best in ensemble, but was quite close to it (difference about 0.5 BLEU).

1 Introduction

As previously noted in (Sennrich et al., 2016; Zhou et al., 2017), the neural machine translation models tend to provide good fluency but sometimes at the expense of the fidelity – they may struggle to cope with rare words, and can exhibit poor coverage/fidelity by ignoring altogether parts of the source.

By training even the same networks on different data one obtains models that have different strengths and weaknesses, sometimes one model provides the better translation, sometimes another one, even if on average they are of rather equal performance.

Our approach, described here, was to automatically select the best translation from a set of candidates produced by an ensemble of neural translators. As the fluency was generally good, as is typically the case with NMT, our heuristic scoring of the translation quality focused on the bidirectional coverage, estimated by making use of a dictionary aided by a set of heuristic rules for the words not found in the dictionary. We aimed to select thus automatically the highest fidelity translation.

Name	Description	Pairs
MED	medication accompanying patient information leaflets from the UFAL Medical Corpus 1.0(ufa) En-Ro(subset)	1048757
NEWS	SE Times En-Ro + Europarl 2017 En-Ro	612422

Table 1: Datasets used to train and validate the neural networks

Combining translators is not new, the most interesting result known to us is (Zhou et al., 2017), where the authors report improvements of over 5 BLEU points in Chinese-to-English translation by combining the outputs of SMT and NMT systems using a neural network.

Our method is much simpler, has the additional advantage of using the NMT models as black-boxes, and requires no further training in addition to that already employed to build the NMT models. The downside is that the BLEU score did not increase over the best in the ensemble (was within 0.5 BLEU of it) on a non directly comparable task, the biomedical field English-to-Romanian translation task of the WMT 2018 workshop.

2 Methods

The datasets listed in Table 1 have been used for training and validation in various ways. We have grouped the En-Ro parallel corpora available to us in two groups, Medical (short: MED) and News+ EU Parliament debates (short: NEWS).

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 644–647

https://doi.org/10.18653/v1/W18-64072

Letter	MED	NEWS
Incorrect ş (unicode 351)	273258	289092
Incorrect ţ (unicode 355)	474633	323086
Correct ș (unicode 537)	28434	101095
Correct ț (unicode 539)	48896	109172

Table 2: Diacritics usage in the datasets used here – number of lines containing a certain letter

2.1 Considerations specific to the Romanian language concerning the character codes used for diacritics

The Romanian language uses 5 letters with diacritics: ă, â, î, ș, ț. Before a 2003 decision of the Romanian Academy, other characters were in wide use instead of ș (unicode 537) and ț (unicode 539): cedilla-based ş (unicode 351) and ţ (unicode 355). The history of decades of broken support in various operating systems and character sets is related at http://kitblog.com/2008/10/romanian_diacritic_marks.html. The diacritics in Romanian are fairly redundant, automatic restoration is possible, with less than 1% errors (Grozea, 2012). The changes over the years, starting with using no diacritics at all in the 1980s and early 1990s, then using cedilla based ones, then comma based ones led to heterogeneous corpora used in NLP: some texts have no diacritics at all, some have the wrong diacritics, some have a mixture of wrong and correct diacritics. This affects multiple NLP tasks, including translation. Learning from examples to translate into Romanian is more difficult than it should be when the examples sampled from various corpora alternate randomly the diacritics they use. The diacritics usage statistics for the datasets used here is given in Table 2.

2.2 NMT models

We have used for our experiments the tensor2tensor (T2T) implementation of the Transformer network (Vaswani et al., 2018). Several training runs have been performed, described in Table 3. The training has been interrupted manually when the loss on the validation set started to increase (early stop), as judged by the experimenter monitoring the evolution of the loss on tensorboard. As such, small fluctuations of the loss do not lead to a too early stop.

The external BPE preprocessing was performed using scripts from the SMT system Moses (Koehn et al., 2007).

2.3 Ensemble Aggregation by Translation Selection

Each model has been used to translate all source sentences from English to Romanian. The aggregation of those outputs has been performed by selecting automatically the translation having the highest quality.

In order to assess the quality of the sentence translations we have computed the percentage of words in the source that have a correspondent in the translation (coverage) and the percentage of the words in the translation that have a correspondent in the source. The minimum of those two numbers between 0 and 1 is taken as the quality of the translation. Once a correspondent is found, it it removed from the next searches (in a greedy fashion, as opposed to the alternative of maximizing the matching with dynamic programming). A word matching is evaluated to 1, when the pair is found in the dictionary, after stemming and the normalization described below, that is applied to the dictionary as well. A pair of words that become identical after stemming and normalization lead to a matching of value 0.3. If the words normalized after stemming are not identical, not too short (they are at least 4 characters) and one of them is a prefix of the other, then the matching is evaluated to 0.2. When computing the coverage mentioned above, the sum of the word pair matching quality is divided by the total number of words.

The preprocessing steps for text normalization, applied both to the sentence pair (source and translation) and on the dictionary are:

- Diacritics removal;
- Replacing of *ph* with *f*, of *y* with *i* and of *ff* with *f*.

The aim of the diacritics removal was to cope with the heterogeneous codes for the letters with diacritics and to cover also for the texts without diacritics. The aim of the substitution of the groups of letters was to increase the chance to recognize proper translation of medical terms originating in Latin or Greek, by bringing them closer to a common phonetic notation.

3 Results

The results are shown in Table 4. The BLEU scores have been computed after replacing the let-

ID	Epochs	Subwords	Train	Validation	Description
1	45000	32768 external BPE	Med	News	early stop, when validation error started to increase
2	28000	32768 external BPE	Med + News	Med	Train on NEWS as well for better fluency
3	35000	32768 external BPE	Med + News	Med	2 trained further
4	28000	16384 T2T subwords	Med	News	repaired diacritics
5	37000	16384 T2T subwords	Med	News	4 trained further
6	48000	32768 T2T subwords	Med	News	like 5, but with larger subwords dictionary

Table 3: Transformer models trained. Models 1-3 used an external Byte Pair Encoding, whereas models 4-6 used the subwords in the tensor2tensor framework to achieve the capability of translating previously unseen words.

ID	BLEU uncased	BLEU cased	Moses
1	20.84	20.54	20.38
2	14.83	14.56	14.38
3	14.10	13.82	13.63
4	**22.48**	**22.16**	**21.99**
5	21.45	21.10	20.90
6	22.12	21.88	21.75
Ensemble	**22.05**	**21.73**	**21.54**

Table 4: BLEU scores evaluated using t2t-bleu from tensor2tensor and multi-bleu-detok from Moses

ters with cedilla-based diacritics both in the translation and in the reference translation with their correct comma-based version.

We have submitted two translations, the one produced by the model with ID=1 in Table 3 (cased BLEU=20.54) and the one produced by the entire ensemble (cased BLEU=21.73).

The run with ID=4 performed best with respect to the BLEU score. The output of the ensemble performed slightly worse than it (by about 0.5 BLEU points), but otherwise being almost equal to the second-best, ID=6.

4 Discussion and Conclusion

We chose to train on the MED corpora and test on NEWS based on the intuition that one can learn from medical texts how to generally translate arbitrary texts, up to the point where excessive specialization on the medical field is detrimental to the performance on the texts in other fields.

There are multiple ways to improve upon this work. The quality of the heuristic depends on the quality of the dictionary, so a straight-forward way would be to use a larger dictionary. The dictionary we have used had approx. 39000 word pairs, but only approx. 17000 Romanian words and approx. 20000 English words; there are multiple pairs for the same source word, when multiple translations exist. For comparison, the Explanatory Dictionary of the Romanian Language (DEX) contains 65000 word definitions.

Another way to improve would be replacing the manually engineered heuristic for evaluating the quality of the translations with one evaluation function learned with machine learning from sentence-aligned parallel corpora. The pair in the training set could then have the label 1 attached to it (with the meaning “correct translation”), whereas variations obtained by eliminating, inserting or changing in a random fashion words from the translation have the label 0 (“incorrect translation”) in the training set.

One reviewer suggested the models could have been combined in the decoder, by combining the word probabilities predictions – we did not try this yet. Each of the 6 members of the ensemble had its own decoder. The advantage in regarding the individual translators as atomic black boxes is that any type of translators can be used, including statistical and human translators. The obvious disadvantage is that in the ideal case the selected translation is the best among the translations to select from, but cannot outperform it; here, it selected reliably one of the best translations.

References

UFAL medical corpus 1.0. https://ufal.mff.cuni.cz/ufal_medical_corpus. Accessed: 2018-07-24.

Cristian Grozea. 2012. Experiments and results with diacritics restoration in romanian. In *International Conference on Text, Speech and Dialogue*, pages 199–206. Springer.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 378–384.