Analyzing Neural MT Search and Model Performance

Jan Niehues, Eunah Cho, Thanh-Le Ha, Alex Waibel
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology, Germany
{jan.niehues,eunah.cho,thanh-le.ha,alex.waibel}@kit.edu

Abstract

In this paper, we offer an in-depth analysis about the modeling and search performance. We address the question if a more complex search algorithm is necessary. Furthermore, we investigate the question if more complex models which might only be applicable during rescoring are promising.

By separating the search space and the modeling using n-best list reranking, we analyze the influence of both parts of an NMT system independently. By comparing differently performing NMT systems, we show that the better translation is already in the search space of the translation systems with less performance. This results indicate that the current search algorithms are sufficient for the NMT systems. Furthermore, we could show that even a relatively small n-best list of 50 hypotheses already contain notably better translations.

1 Introduction

Recent advances in NMT systems (Bahdanau et al., 2014; Cho et al., 2014) have shown impressive results in improving machine translation tasks. Not only it performed greatly in recent machine translation campaigns (Cettolo et al., 2015; Bojar et al., 2016) measured in BLEU (Papineni et al., 2002), it is considered to be able to generate sentences with better fluency.

Despite the successful results in translation performance, however, the optimality of the search algorithm in NMT has been left under-explored. In this work, we analyze the influence of search and modeling of an NMT system by evaluating them separately. We aim to demonstrate whether further research on the model development is more promising or the one on the search algorithm would be more beneficial.

We attempt to simulate this by n-best rescoring using different models. For this, n-best lists are rescored by different models including the one which generated them. Additionally we build a configuration with all n-best lists joined, in order to see whether rescoring this joined n-best list using the same model would bring a performance boost.

2 Related Work

There has been a number of works devoted to combine different systems from the same or different machine translation (MT) paradigms using n-best lists of hypotheses (Matusov et al., 2006; Heafield et al., 2009; Macherey and Och, 2007). The hypotheses are aligned, combined and scored by a model to produce the best candidate according to a metric. There was a thorough analysis on how the size of n, the diversity of the outputs from different systems and performance of individual systems can affect the final translation of the system combination. Hildebrand and Vogel (2008) examine the feature impact and the n-best list size of such a combination of phrase-based, hierarchical and example-based systems. Gimpel et al. (2013) show how diversity of the outputs and the size of the n-best lists determine the performance of the combined system.

Costa-Jussà et al. (2007) analyze the impact of the beam size used in statistical machine translation (SMT) systems. Wisniewski and Yvon (2013) conduct an in-depth analysis over several types of errors. Based on their proposal to effectively calculate oracle BLEU score for an SMT system, they can separate the errors due to the restriction of the

Proceedings of the First Workshop on Neural Machine Translation, pages 11–17,
Vancouver, Canada, August 4, 2017.

search space (search error) from the errors due to models not good enough to cover the best translation (model error). Although this work is the closest to our work in terms of analysis methods, our work differs from theirs by addressing the issue focused on the NMT systems.

In Neubig et al. (2015), the size of the n-best list produced by a phrase-based SMT and rescored by an NMT is taken into account for an error investigation. The work also shows which types of errors from the phrase-based system can be corrected or improved after NMT rescoring. To the best of our knowledge, our work is the first to examine the impact of search and model performance in pure NMT systems.

2.1 Neural Machine Translation

Neural machine translation, whilst considered to be in the same direction with phrase-based SMT from a statistical perspective, is actually separable from traditional SMT in terms of how it models the representation of source and target sentences as well as the translation relationship between them. In this section, we describe the general architecture of a NMT system in order to understand the needs and importance of such an analysis. The NMT architecture described here is similar to the attention-based NMT from Bahdanau et al. (2014).

An attentional NMT system consists of an encoder representing a source sentence and an attention-aware decoder that produces the translated sentence.

The encoder which is comprised of bidirectional recurrent layers reads words from the source sentence and encodes them into annotation vectors. Each annotation vector contains the information of the source sentence related to the corresponding word from both forward and backward directions.

A single layer featuring attention mechanism allows the decoder to decide which source words should take part in the prediction process of the current target word. Basically, attention layer examines a context vector of the source sentence which is weighted sum of all annotation vectors and normalized, where the weights reflect some relevance between previous target words and all the source words.

The decoder, which is also recurrent-based, recursively generates the target candidates with their probabilities to be selected based on the context vector from the attention layer, the previous recurrent state and the embedding of the previously chosen word.

The whole network is then trained in an end-to-end fashion to learn parameters which maximizes the likelihood between the outputs and the references. In the testing phase, a beam search is utilized to find the most probable target sequences giving the n-best list from the architecture.

We could see that in NMT, therefore, the model (e.g. the ways the encoder representing a source sentence or the attentional layer modeling attention mechanism) and the search algorithm are one of the most important aspects to be analyzed.

3 Search and Model Performance

In this analysis we evaluate the search and modeling performance of NMT. In order to evaluate them individually, we need to separate the modeling errors and the search errors of the system. While the search in phrase-based MT was relatively complex, the search algorithm in NMT is relatively straightforward. In state-of-the-art system, a beam search algorithm is used with a small beam between $10 - 50$.

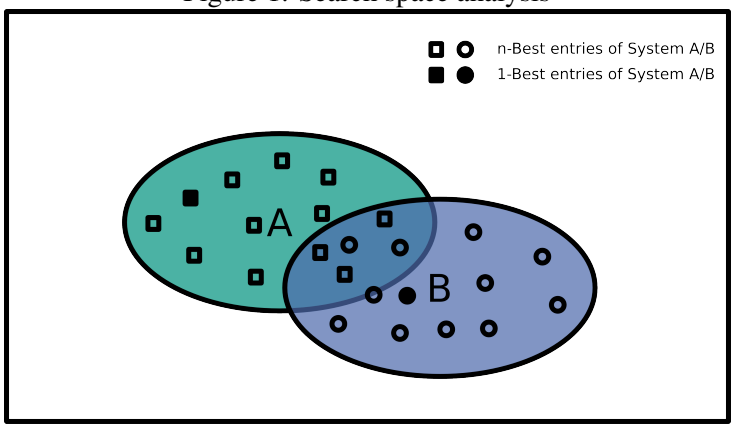
The goal of this work is to establish whether improvements on the NMT model itself is more promising or the ones on the search algorithm. If there are many search errors due to the pruning during decoding, a better search algorithm would be promising. In contrast, if there are relatively few search errors, further research on the model is more promising.

3.1 Analysis Setup

A straightforward way would be to evaluate all possible hypotheses. In this case we do not have any search error and can directly measure the modeling errors. However this cannot be performed efficiently since the number of all possible hypotheses is very large. Therefore, we analyzed the performance of two or several systems with different performances.

In the experiments, for example, we have systems A and B where the translation performance of A is better than the one of B. Then we approximated the search space of A and B by their n-best lists and evaluated the performance of each system in the search space of A by scoring the n-best list with the model and selecting the hypothesis with

Figure 1: Search space analysis



the highest probability. Figure 1 shows the search spaces of system A and B, approximated by their n-best hypotheses. Their 1-best entries are also marked accordingly.

The question we address is why the system B did select its best hypothesis ● and did not select the better-performant hypothesis ■. One reason might be that ■ is not in the system B's search space and therefore the system could not find it. The other reason might be that system B prefers ● over ■. In this case, we need to improve the modelling.

If the performance of model B on the n-best list of A is better than the initial score of B, it suggests that the model B is able to select a better hypothesis and therefore the search is not optimal. On the other hand, if the performance is similar, it means that B is not able to select a better hypothesis, even though there are better ones according to the evaluation metric.

In the experiments, we used two different ways of constructing the models A and B. In a first series of experiments, we used the single best system as well as ensemble systems. In a second series, we used systems using different ways to generate the translations. Details of the systems will be given in the Section 4.

4 System Description

Our German↔English NMT systems are built using an encoder-decoder framework with attention mechanism, `nematus`.[1] Byte pair encoding (BPE) is used in order to generate sub-word units (Sennrich et al., 2015). Long sentences whose sentence length exceeds 50 words are exempted from the training. We use minibatch size 80 and sentences are shuffled within every minibatch. Word embedding of size 500 is applied, with hidden layers of size 1024. Dropout is applied at every layer with the probability 0.2 in the embedding and hidden layers and 0.1 in the input and output layers. Our models are trained with Adadelta (Zeiler, 2012) and the gradient norm is clipped to 1.0. For the single models, we apply the early stopping based on the validation score.

The baseline system is trained on the WMT parallel data, namely EPPS, NC, CommonCrawl and TED corpus. As validation data we used the newstest13 set from IWSLT evaluation campaign. Therefore, this data is from TED talks. Test is applied on two domains. First domain is TED talks, same as the optimization set. We use newstest14 for this testing. Another domain is telephone conversation and we used MSLT (Christian Federmann, 2016) for testing. Since no exact genre-matching development data is published for the evaluation campaign (Cettolo et al., 2015), we used the TED-optimized system for the MSLT testing. For each experiment, we also offer oracle BLEU scores on the n-best lists, calculated using multeval (Clark et al., 2011).

4.1 Configurations

We tried different system configurations to generate and rescore the n-best lists. By using 40K operations of BPE we had *SmallVoc* configuration, and with 80K *BigVoc* configuration. In *SmallVoc.rev*, target sentence are generated in the reversed order. In *SmallVoc.mix*, target side corpus is joined with the source side corpus to form a mixed input as described in Cho et al. (2016). We build an NMT system which takes pre-translation

[1] `https://github.com/rsennrich/nematus`

from a PBMT system following the work in Niehues et al. (2016), which will be referred as *PrePBMT*. A configuration using more monolingual data for this training is called *PrePBMT.large*. In *Union*, we use the joined n-best lists from different systems.

4.2 n-best list

All n-best lists are generated for $n = 50$ from a standard beam search. The size of n-best lists are limited due to time and computational limitation. In our preliminary experiments where we increased n-best list from 1 to 50, it did not significantly change the performance of one model. Therefore, in this work, we approximate the 50-best lists our search space and conducted the analysis. By doing so, we also aim to give a practical analysis on a model vs. search performance comparison in NMT and useful guidelines from it.

5 Analysis on the Results

In this section, we discuss the experimental results and detailed analysis. In the first part, we discuss the results of the experiments on the baseline systems. In the second part, we combine NMT systems that use different text representations.

5.1 NMT Baseline Systems

In this section, we analyze the performance of baseline systems. It largely breaks down to two tasks: TED and MSLT translation.

5.1.1 TED translation

Table 1 shows the baseline system performance on the TED translation task, from German to English. The table is showing translation performance of reranking each n-best list using different models.

Model / n-best	Single	Ensemble	Oracle
Single	31.96	32.37	41.81
Ensemble	32.09	32.41	42.31
Union	31.95	32.39	44.55

Table 1: Baseline: TED German→English

For the *Single* system, we took the best-performant *BigVoc* system. The *Ensemble* system is then generated by combining several training steps of the single system training. The *Union* n-best list is the joined n-best list of all the individual systems used in the ensemble. For building a *Ensemble* system, we combine different models from several time steps of *Single* training. Then in the softmax operation, normalized probabilities of each word are considered. As mentioned earlier, we also offer the oracle BLEU scores given each n-best list.

Model performance As shown in the table, we can improve the translation performance by 0.5 BLEU point by using the *Ensemble* system to rescore n-best list generated by the same system, compared to the same case for *Single* system. The main contribution for this improvement seems to be the better modeling. When we use the *Single* model to rescore the *Ensemble* or *Union* n-best list, we get mainly the same performance. Thus, the reason for the relatively lower performance of the *Single* system is considered to be that it does not model the translation probabilities better, not because it does not find better translations. The oracle scores indicate the similar trend. When the n-best list is large (*Union* setup), we have better translations in the n-best list. However, these hypotheses were not selected by either of the models.

Search performance The numbers in Table 1 suggests that the search is well-performant in NMT. For example, when we use the *Ensemble* model to rescore the *Single* n-best list, the translation performance reaches 32.37 BLEU points. At the same time, when we use the same model to rescore the *Ensemble* n-best list, we achieve a similar performance.

5.1.2 MSLT translation

The performance on the single system on the MSLT task (Christian Federmann, 2016) is shown in Table 2.

Model / n-best	Single	Ensemble	Oracle
Single	34.63	38.35	53.85
Ensemble	35.94	38.80	56.46

Table 2: Baseline: MSLT German→English

In this task, rescoring *Single* n-best list using the same model itself performs around 4 BLEU worse than rescoring *Ensemble* n-best list using the *Ensemble* model. Also, we can observe that the *Single* model performs better when using the n-best list of the *Ensemble* model.

We find two explanations for this improvement. A) The *Ensemble* n-best list contains better-

performing hypotheses that the *Single* model did not find during the search. Or alternatively, B) the *Ensemble* n-best list does not contain the hypotheses that are good according to the *Single* model but not according to the evaluation metric. In this case, the model would select different hypotheses.

In order to locate search error, we evaluated and compared the model score of hypothesis chosen from different n-best lists. Only in 2.5% of the chosen hypotheses, the score of the hypothesis selected from the *Ensemble* n-best list is higher than the one from the *Single* n-best list. Thus, we have a search error only in these cases.

In contrast, in 90.7% of the sentences, the score from the *Single* n-best list is higher. The main reason for the improvement, therefore, is not considered to be better search. Rather, the search space by the *Ensemble* system does not contain the worse-performing translations which are highly ranked by the *Single* system.

The n-best lists of the *Single* model contains well-performing translations. For example, the performance achieved when using the *Ensemble* model to rescore the *Single* n-best list is almost similar to the one achieved when applying the same model on the *Ensemble* n-best list. This performance is nearly 4 BLEU points better than rescoring the same n-best list using the *Single* model.

While the performance of the *Ensemble* model on both n-best lists is similar, interestingly, the oracle score of the *Ensemble* n-best list is clearly higher. Therefore, the models seem not able to select better translations in the *Ensemble* n-best list compared to the *Single* n-best list.

5.2 NMT Text Representation Systems

As a next line of experiment, we combine NMT systems that use different text representations.

5.2.1 TED translation

Table 3 lists the systems used in the experiment and their performance on the TED task.

We can observe that the results of *Union* rescored by each model is similar to the performance of the model's n-best list rescoring, as marked in bold letters in each column. Considering that the *Union* n-best list is considerably larger, it seems again that the model can find the best hypothesis according to the model.

In contrast, if we use all models (*All*) by using sum of log probabilities of all models to rescore the n-best lists, we achieve similar performance for all n-best lists. Thus, it seems that all 50-best lists contain already very good hypotheses. Only the n-best list of the *PrePBMT* system seems to contain relatively worse options. This is also shown by the oracle scores. One reason could be that the pre-translation by the PBMT system is guiding the search and therefore the n-best list contains relatively limited variety.

In addition, we observe that the performance of each model on its own n-best list is considerably worse than the model rescoring other n-best lists. This can be explained by the following phenomena: some translations of a system A are highly-ranked by the model itself, but not by the others. Therefore, they are selected by the system A but not in the n-best lists of the other systems. If they are in the n-best list, e.g. in the n-best lists of the system A and in the *Union*, they will be selected only when using the system A, leading to worse performance in BLEU. In contrast, if we use different n-best lists, the translation performance is better.

English→German In addition, we extend this experiment to another language direction. Table 4 shows the results when the same experiment is applied to En-De TED task.

Here the same phenomena is observed. Again, the *Union* n-best list does not improve the translation quality. Nonetheless, the oracle score is significantly higher indicating that the model finds the better hypotheses. Furthermore, the n-best lists already contain better hypotheses which can be chosen using better models, i.e. the combination of all models.

5.2.2 MSLT translation

Table 5 shows the similar results when the same experiments are applied to the MSLT task. The *Union* configuration performs similar to rescoring using the same model, while performing considerably worse than the case where the same n-best list rescored by other models.

6 Conclusion

Our experiments on two language pairs and two different tasks showed that there are only few search errors in the state-of-the-art NMT systems. Even when better hypotheses are added in the n-best list, the models do not select a different hypothesis. Thus, the search algorithms seem to be

n-best list \ Model	SmallVoc	SmallVoc.rev	BigVoc	PrePBMT	All	Oracle
SmallVoc	**31.74**	32.17	32.62	32.55	33.03	41.82
SmallVoc.rev	32.24	**31.28**	32.58	32.06	32.93	40.97
BigVoc	32.57	32.50	**32.41**	32.63	33.26	42.31
PrePBMT	32.19	31.97	32.53	**31.41**	32.65	40.67
Union	**31.83**	**31.27**	**32.42**	**31.39**	33.24	46.58

Table 3: Text representation systems: TED German→English

n-best list \ Model	SmallVoc.mix	BigVoc	PrePBMT	PrePBMT.large	All	Oracle
SmallVoc.mix	**26.19**	27.09	26.93	27.03	27.12	33.71
BigVoc	26.97	**27.28**	27.26	27.12	27.48	34.16
PrePBMT	26.96	27.00	**26.44**	27.15	27.14	32.95
PrePBMT.large	27.25	27.47	26.85	**27.03**	27.41	33.78
Union	**26.25**	**27.28**	**26.44**	**27.03**	27.76	38.95

Table 4: Text representation systems: TED English→German

sufficient.

Furthermore, we showed that a relatively small n-best list of 50 entries already contains notably better translation hypotheses. This result indicates that improving rescoring models are promising for performance boost. In this work, we showed that it is often sufficient to use a model in rescoring only. This finding also motivates the development of models which are challenging to use directly during the decoding, such as bi-directional decoders.

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452. The research by Thanh-Le Ha was supported by Ministry of Science, Research and the Arts Baden-Württemberg. This work was supported by the Carl-Zeiss-Stiftung.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT16)*. Association for Computational Linguistics, Berlin, Germany, pages 12–58.

M Cettolo, J Niehues, S Stüker, L Bentivogli, and M Federico. 2015. The IWSLT 2016 Evaluation Campaign. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2015)*. Seattle, US.

Eunah Cho, Jan Niehues, Thanh-Le Le, Matthias Sperber, Mohammed Mediani, and Alex Waibel. 2016. Adaptation and combination of nmt systems: The kit translation systems for iwslt 2016. In *IWSLT*. Seattle, WA, USA.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Doha, Qatar.

William D. Lewis Christian Federmann. 2016. Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *IWSLT*. Seattle, WA, USA.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 176–181. http://dl.acm.org/citation.cfm?id=2002736.2002774.

Model / n-best list	SmallVoc	SmallVoc.rev	BigVoc	PrePBMT	All	Oracle
SmallVoc	**37.90**	39.50	39.35	39.05	40.30	56.41
SmallVoc.rev	39.74	**38.72**	39.92	39.94	40.80	56.82
BigVoc	38.73	39.61	**38.80**	39.51	40.25	56.46
PrePBMT	38.91	39.68	39.36	**38.33**	40.24	54.44
Union	**37.92**	**38.65**	**38.81**	**38.33**	40.66	63.09

Table 5: Text representation systems: MSLT German→English

Marta R Costa-Jussà, Josep M Crego, David Vilar, José AR Fonollosa, José B Mariño, and Hermann Ney. 2007. Analysis and System Combination of Phrase-and N-gram-based Statistical Machine Translation Systems. In *The Conference of the North American Chapter of the Association for Computational Linguistics (NAACLT'07)*. Association for Computational Linguistics, Rochester, NY, USA, pages 137–140.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A Systematic Exploration of Diversity in Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. Seattle, WA, USA.

Kenneth Heafield, Greg Hanneman, and Alon Lavie. 2009. Machine translation system combination with flexible word ordering. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT'09)*. Association for Computational Linguistics, Athens, Greece, pages 56–60.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of Machine Translation Systems via Hypothesis Selection from Combined N-Best Lists. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA'08)*. Hawaii, USA, pages 254–261.

Wolfgang Macherey and Franz J Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems .

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing Consensus Translation for Multiple Machine Translation Systems Using Enhanced Hypothesis Alignment. In *Proceedings of the 11th Conference of the European Chapter of Association for Computational Linguistics (EACL 2006)*. Association for Computational Linguistics, Trento, Italy, pages 56–60.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT'15)*. Kyoto, Japan.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *the 26th International Conference on Computational Linguistics (Coling 2016)*.

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 (ACL 2002)*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany.

Guillaume Wisniewski and François Yvon. 2013. Oracle Decoding as a New Way to Analyze Phrase-based Machine Translation. *Machine Translation* 27(2):115–138.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. In *CoRR*.