Exploring Hypotheses Spaces in Neural Machine Translation

Frédéric Blain f.blain@sheffield.ac.uk
Lucia Specia l.specia@sheffield.ac.uk
Pranava Madhyastha p.madhyastha@sheffield.ac.uk
Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK

Abstract

Both statistical (SMT) and neural (NMT) approaches to machine translation (MT) explore large search spaces to produce and score translations. It is however well known that often the top hypothesis as scored by such approaches may not be the best overall translation among those that can be produced. Previous work on SMT has extensively explored re-ranking strategies in attempts to find the best possible translation. In this paper, we focus on NMT and provide an in-depth investigation to explore the influence of beam sizes on information content and translation quality. We gather new insights using oracle experiments on the efficacy of exploiting larger beams and propose a simple, yet novel consensus-based, n-best re-ranking approach that makes use of different automatic evaluation metrics to measure consensus in n-best lists. Our results reveal that NMT is able to cover more of the information content of the references compared to SMT and that this leads to better re-ranked translations (according to human evaluation). We further show that the MT evaluation metric used for the consensus-based re-ranking plays a major role, with character-based metrics performing better than BLEU.

1 Introduction

There has a been a recent surge of interest and work in the field of end-to-end, encoder-decoder neural machine translation (NMT). In the last two years, such approaches surpassed the state-of-the-art results by the then *de facto* statistical machine translation approaches (SMT) (Bojar et al., 2016a). While NMT systems are trained end-to-end using as a single model, SMT systems use a pipeline-based approach that make use of several components. This means that NMT systems are jointly optimised for both better encoding and better decoding. SMT systems, on the other hand, decompose the problem by first finding plausible sub-sentence translation candidates given some training data, such as phrases in phrase-based SMT (Koehn et al., 2003), and then scoring such candidates utilising components such as the translation and language models. Both types of systems are markedly different in their approaches to transform source into target language and in the information they explore.

Given a source sentence, at decoding time both types of approaches can explore hypotheses spaces to pick the best possible translation. Most of current implementations of both statistical and neural MT approaches use beam search for that. It has been observed that NMT systems, when compared to their statistical counterparts, use smaller beam sizes, and yet are able to obtain better translations for the same source sentences (Bahdanau et al., 2014; Stahlberg et al., 2017). Smaller beam sizes boost the speed of decoders (Luong et al., 2015; Bahdanau et al., 2014). In addition, it has been reported (Stahlberg et al., 2016) that neural approaches do not significantly benefit from large beam sizes. In fact, beam sizes of 8–12 are the most common in NMT. Statistical approaches, on the other hand, usually search over larger beam sizes (of orders of 100s) (Lopez, 2008).

There have been multiple approaches proposed in the context of SMT that explore the n-best generated translation hypotheses using beam search (Och et al., 2004; Shen et al., 2004; Lambert and Banchs, 2006; Hasan et al., 2007; Duh and Kirchhoff, 2008). Since models used for scoring translation hypotheses and metrics used to evaluate the final translation quality are different, one of the strategies is to learn a re-ranking model for n-best hypotheses based on the evaluation metric of interest. We further detail this and other strategies in Section 2. However, to the best of our knowledge, there is little research that systematically looks at the effect of beam sizes or explores n-best hypotheses in the context of NMT.

We summarise our contributions in this paper as follows: (a) We investigate the influence of beam size on the search space, as well as on the information content of translations (Section 4); and (b) We present a new re-scoring approach for n-best re-ranking based on information overlap amongst MT candidates within the n-best list according to different automatic MT evaluation metrics. We report results that include human evaluation to assess the quality of alternative translations produced by this approach versus baseline systems (Section 5). We observe that our approach leads to better translation choices. We also observe that in most cases the best translation hypothesis is chosen among those generated from using larger beam sizes. These results are based on four language pairs and different datasets and evaluation metrics (Section 3).

2 Background

In what follows, we briefly describe background on the decoding process in SMT and NMT approaches, as well as related work on exploring n-best lists for improved translation quality.

Beam search decoding in SMT In SMT decoding, the standard procedure is to perform the search for the best translation given the (often pruned) space of possible translations based on a combination of the scores estimated for its model components, each component capturing a different aspect of translation (word order, translation probability, etc.). This is done through a heuristic method using stack-based beam search. In phrase-based SMT (Koehn et al., 2003), given a source sentence, the decoder fetches phrase translations available in the phrase table and builds a graph starting with an initial state where no source words have been translated and no target words have been generated. New states are created in the graph by extending the target output with a phrase translation that covers some of the source words not yet translated. At every expansion, the current cost of the new state is the cost of the original state multiplied with the model components under consideration. Final states in the search graph are hypotheses that cover all source words. Among these, the hypothesis with the lowest cost (highest model score) is selected as the best translation. Often a threshold is used to define a *beam* of good hypotheses and prune the hypotheses that fall out of this beam. The beam follows the (presumably) best hypothesis path, but with a certain width to allow the retention of comparable hypotheses, i.e. neighbouring hypotheses that are close in score from the best one (Koehn, 2010).

If an exhaustive search was to be performed, then all translation options, in different orders, could be used to build alternative hypotheses. However, in practice the search space is pruned in different ways and only the most promising hypotheses are kept, with early pruning potentially eliminating good hypotheses from the search space. In principle, larger beams would thus allow for more variation in the n-best lists, while potentially introducing lower quality candidates, but also giving seemingly *bad* candidates a chance to obtain higher scores in later stages of decoding. There is therefore a direct relationship between the size of the beam and the maximum number of candidates that can be generated in the n-best list. However, the actual candidates in the n-best list are also affected by other design choices, such as the pruning and hypotheses combination strategies used (Lambert and Banchs, 2006; Duh and Kirchhoff, 2008; Hasan et al., 2007).

In addition, different approaches have been proposed to specifically promote diverse translations in SMT systems' n-best lists. These include using compact representations like lattices and hypergraphs (Tromble et al., 2008; Kumar and Byrne, 2004) and establishing explicit conditions during decoding. Gimpel et al. (2013), for example, add a dissimilarity function based on n-gram overlaps, choosing translations that have high model scores but are distinct from already-generated ones.

Beam search decoding in NMT NMT decoding also relies on beam search, but the process is much more expensive than in SMT and thus a limited beam size is often used, leading to narrow hypotheses spaces (Li and Jurafsky, 2016; Vijayakumar et al., 2016). Given a certain prespecified beam size k, k-best lists are generated in a greedy left-right fashion retaining only the top-k candidates as follows: at the first time step in decoding, a fixed-number k hypotheses are retained based on the highest log-probability (model score) of each generated word. Each of the k hypotheses is expanded at each time-step by selecting top k word translations. This continues until the end-of-sequence symbol is obtained. The highest scoring candidate is retained and stored into the final candidate list followed by a decrease of beam by one. The whole process continues until the beam is reduced to zero. Finally, the best translation hypothesis amongst the list is the one with highest log-probability. We note here that in most NMT approaches both the set of hypotheses and the beam size are equivalent. Essentially, the NMT decoder obtains the top translation hypotheses that maximise the conditional probability given by the model.

Li and Jurafsky (2016) increase diversity in the n-best list by adding an additional component to the score used by the decoder to rank k hypotheses at each time step. This component rewards top-ranked hypotheses generated from each ancestor, instead of ranking all candidates from all ancestors together. Similarly, Vijayakumar et al. (2016) propose *Diverse Beam Search*, where they optimise an objective with two terms: the standard cross entropy loss and a dissimilarity term that encourages beams across groups to differ.

N-best re-ranking in SMT In addition to having access to only a subset of the search space, the model components used in SMT only provide an estimate of translation quality. As a consequence, using only the hypothesis ranked as the best by the decoder often leads to suboptimal results (Wisniewski et al., 2010; Sokolov et al., 2012a). Therefore, it is common practice in SMT to explore other hypotheses in the search space, the so called *n-best list*. Re-ranking an n-best list of candidates produced by an SMT system has been a long standing practice. The general motivation for doing so is the ability to use additional information in the process, which is unavailable or too costly to compute at decoding time, e.g. syntactic features of the entire sentence (Och et al., 2004), estimates on overall sentence translation quality (Blatz et al., 2003), word sense disambiguation scores (Specia et al., 2008), large language model scores (Zhang et al., 2006), and translation probability from a neural MT model (Neubig et al., 2015), among others.

This additional information is usually treated as new model components and combined with the existing ones. Various techniques have been proposed to perform n-best list re-ranking. They generally learn weights to combine the new and existing model components using algorithms such as MIRA (Crammer and Singer, 2003) with linear[1] or non-linear functions (Sokolov et al., 2012b), as well as more advanced methods, such as multi-task learning (Duh et al., 2010). Hasan et al. (2007) provides a study on the potential improvements on final translation quality by exploring n-best lists of different sizes. They show that even though oracle-based re-ranking

[1] `https://github.com/moses-smt/mosesdecoder/tree/master/scripts/nbest-rescore`

on very large (100,000 hypotheses) n-best lists yields the best translation quality, automatic re-ranking methods reach a plateau on the improvement after 1,000 hypotheses. Very large n-best lists will contain very many noisy translations, so they suggest that only with extremely accurate re-ranking methods one should explore such large spaces.

In an attempt to have a more reliable way to score translation candidates, Kumar and Byrne (2004) introduced the Minimum Bayes Risk (MBR) decoding approach and used it to re-rank n-best hypotheses such that the best hypothesis is the one that minimises the Bayes-risk defined in terms of the model score (translation probability) and a loss function computed between the translation hypothesis and a gold translation (e.g. a translation quality metric such as BLEU (Papineni et al., 2002)). This method has been shown to be beneficial for many translation tasks (Ehling et al., 2007; Tromble et al., 2008; Blackwood et al., 2010). They have however only experimented a fixed n (1,000).

N-best re-ranking in NMT While there is a large body of literature that investigates different strategies for exploring n-best hypotheses spaces in SMT, there have been very few attempts at exploring such spaces in NMT. Stahlberg et al. (2017) adapt MBR decoding to the context of NMT and to be used for partial hypotheses rather than entire translations. The NMT score is combined with the Bayes-risk of the translation according to the SMT lattice. This approach goes beyond re-scoring of n-best lists or lattices as the neural decoder is not restricted to the SMT search space. The resulting MBR decoder produces new hypotheses that are different from those in the SMT search space.

Li and Jurafsky (2016) propose an alternative objective function for NMT that maximises the mutual information between the source and target sentences. They implement the model with a simple re-ranking method. This is equivalent to linearly combining the probability of the target given the source, and vice-versa. An NMT model is trained for each translation direction, and the source→target model is used to generate n-best lists. These are then re-ranked using the score from the target→source model. Shu and Nakayama (2017) studies the effect of beam size in NMT MBR decoding. They considered beams of size 5, 20 and 100 and found that while in standard decoding increasing the beam size is not beneficial, MBR re-ranking is more effective with a large beam size.

Comparison between NMT and SMT There has been increasing interest in systematically studying differences between NMT and SMT approaches. Bentivogli et al. (2016) conducted an analysis for English→German translations by both NMT and SMT systems. They conclude that the outputs of the NMT system are better suited in terms of syntax and semantics, with better word order and less human post-editing effort required to fix the translations. They observe that the average sentence length in an SMT system is always longer than in an NMT system. This could be attributed to the optimisation of the cross-entropy loss and the fact that the outputs are chosen on the basis of the log-probability scores in NMT systems.

Toral and Sánchez-Cartagena (2017) conducted an in-depth analysis on a set of nine language pairs to contrast the differences between SMT and NMT systems. They observe that the outputs of NMT systems are more fluent and have better word order when compared to SMT systems. They note that despite the smaller beam sizes in NMT in general the top outputs of the NMT system for a given source sentence are more distinct than the top outputs from SMT systems. However, it is not clear whether or not they explore distinct n-best options from the SMT or a mixture of distinct and non-distinct options. Both previous studies conclude that the NMT systems perform poorly when translating very long sentences.

3 Experimental Settings

In this section we describe the data, tools, metrics and settings used in our experiments to investigate the influence of beam size in the generated translations.

Language Pairs We report results with NMT systems – the focus of this paper – for four language pairs: English↔German and English↔Czech. For English↔Czech we also report results with SMT systems for comparison.

NMT Systems We use the freely available Nematus (Sennrich et al., 2016) toolkit and its pre-trained models[2] for English↔German and English↔Czech. The Nematus systems are based on attentional encoder-decoder neural machine translation approach (Bahdanau et al., 2014) and were built after *Byte-Pair Encoding* (Sennrich et al., 2015b).[3] The models were trained as described in (Sennrich et al., 2016) using both parallel and synthetic (Sennrich et al., 2015a) data under the constrained variant of the WMT16 MT shared task, mini batches of size 80, a maximum sentence length of 50, word-embeddings of size 500, a hidden layers of size 1024, and Adadelta as optimiser (Zeiler, 2012), reshuffling the training corpus between epochs.These models were chosen as they have been highly ranked in the evaluation campaign of the WMT16 Conference (Bojar et al., 2016c).

SMT Systems We use pre-trained models from the Tuning shared task of WMT16 for English↔Czech to build SMT systems for comparison. These models were built using the Moses toolkit (Koehn et al., 2007) trained on the CzEng1.6pre[4], (Bojar et al., 2016b) a 51M parallel sentences corpus built from eight different sources. The data was tokenised using Moses tokeniser (Koehn et al., 2007) and lowercased; sentences longer than 60 words and shorter than 4 words were removed before training. The weights were determined as the average over three optimisation runs using MIRA (Crammer and Singer, 2003) towards BLEU. Word alignment was done using fast-align (Dyer et al., 2013) and for all other steps the standard Moses pipeline was used for model building and decoding. This was reported as the best system for English↔Czech (Jawaid et al., 2016).

By using pre-trained and freely available models for our NMT and SMT systems, we have consistent models amongst the different language pairs and results can be more easily reproducible.

Beam Settings SMT systems usually employ a large beam. In the training pipeline of the Moses decoder, the beam size is set by default to 200. NMT systems, on the other hand, normally use a much smaller beam size of 8 to 12. This is assumed to offer a good trade off between quality and computational complexity. We note that the implementations of n-best decoding is different in both NMT and SMT. In most NMT systems, there is a 1-to-1 correspondence between the beam size and the n-best list size. Therefore, we will use the term n*-best* to refer to the output of an NMT system with a beam of size n, and to the n best outputs of an SMT system, where the beam size has been set, by default, to 200.

We also note that the translations in the n-best list produced by NMT are always different from each other, even though only marginally in many cases (e.g. a single token). In SMT, one can choose whether or not only distinct candidates should be considered. We report on distinct options only to gather insights on the diversity in n-best lists in SMT versus NMT.

Metrics For our experiments we consider three automatic evaluation metrics amongst the most widely used and which have been shown to correlate well with human judgements (Bojar

[2]http://data.statmt.org/rsennrich/wmt16_systems/

[3]The models were obtained from http://statmt.org/rsennrich/wmt16_systems/

[4]http://ufal.mff.cuni.cz/czeng/czeng16pre

et al., 2016c): **BLEU**, an n-gram-based precision metric which works similarly to position-independent word error rate, but considers matches of larger n-grams with the reference translation; **BEER** (Stanojevic and Sima'an, 2014), a trained evaluation metric with a linear model that combines features capturing character n-grams and permutation trees; and **ChrF** (Popovic, 2015), which computes the F-score of character n-grams. These metrics are used both for evaluating final translation quality and for measuring similarity among translations in our consensus-based re-ranking approach.

4 Effect of Beam Size

Current work in NMT takes a beam size of around 10 to be the optimal setting (Sennrich et al., 2016). We empirically evaluate the effect of increasing the beam size in NMT to explore n-best of sizes 10, 100 and 500. The goals are to understand (a) the informativeness of the translations produced; (b) the scope for obtaining better translations by simply exploiting the n-best candidates, similarly to previous work in SMT.

4.1 Effect of Beam Size on Information Content of Translations

We define information content as the word overlap rate between the system generated translation and the reference translation. We further break this into two categories:

1. *% covered*: This indicates the average proportion of words that are shared between the (a) 1-best output of the MT system and the reference translation, or (b) all the n-best outputs and the reference translation. It is computed by looking at the intersection between the vocabulary of the MT candidate(s) and the one of the reference, averaged at corpus-level.

2. *% exact match*: This indicates the proportion of sentences that are exact matches between (a) the 1-best of the MT system and the reference translation, and (b) all the n-best outputs and the reference translation.

This is similar to the approach in (Lala et al., 2017) where the authors measure word overlap with respect to system outputs, but their focus is on multimodal NMT. *% covered* approximates indicates the word-level precision of the MT system, given the n or 1-best candidates and the reference translation, and *% exact match* approximately indicates the sentence-level recall given the n or 1-best candidates and the reference translation.

Our intuition here is that if the systems are adequately trained, increasing the beam size – and thereby the n-best list length – should result in obtaining a larger word overlap with reference translation, and potentially a larger number of exact matches at the sentence level, although the latter is a much taller order. We note that since only one reference translation is available, mismatches between words in the MT output and reference translations could reflect acceptable variances in translation.

Observations and Discussion In Table 1 we report the scores of each MT system using BLEU, BEER and ChrF3 on the WMT16 test sets with different sizes of n-best lists: for NMT we report sizes 10, 100 and 500, while for SMT we report a 500-best list with a beam size set to the default size of 200. Since there is no 1-to-1 relationship between beam sizes and n-best list sizes in SMT, reporting on different beam sizes would require arbitrarily choosing a specific n for each beam size. We instead chose the largest n also used for the NMT experiments (500), and a large enough beam size (200). The metric scores are computed on the 1-best translation, which may vary if different beam sizes are used. We observe that for NMT increasing the n-best size from 10 to 100 helps improve the performances for English↔German translations. For English↔Czech, we do not observe any gain, but rather a significant drop. Also, if the beam size is too large (500 in our case), the performance drops for all language pairs. This indicates that larger beam sizes do not necessarily lead to better 1-best translations, and that the choice can be a function of the language pair and the dataset. This seems to suggest that with such large beam sizes many translation candidates, including spurious ones, end up being ranked as the 1-best, most likely because of limitations in the functions used to score translation candidates.

NEURAL MT	**English→German**			**German→English**		
n-best	BLEU	BEER	ChrF3	BLEU	BEER	ChrF3
n=10	26.73	60.20	59.20	32.58	61.84	60.61
n=100	26.82	60.25	59.33	32.68	61.91	60.74
n=500	26.18	60.12	59.12	32.70	61.91	60.75
	English→Czech			**Czech→English**		
n-best	BLEU	BEER	ChrF3	BLEU	BEER	ChrF3
n=10	18.50	53.90	51.45	26.26	58.03	56.00
n=100	18.31	53.83	51.37	26.17	58.00	56.00
n=500	17.81	53.67	51.25	24.19	57.57	55.62

STATISTICAL MT	**English→Czech**			**Czech→English**		
n-best	BLEU	BEER	ChrF3	BLEU	BEER	ChrF3
n=10/100/500	10.64	48.88	46.51	18.19	52.59	51.32

Table 1: Translation quality results on the WMT16 test sets for both NMT and SMT systems using n-best lists of sizes 10, 100 and 500. The scores are computed on the 1-best translation towards the reference translation.

In Table 2 we report our empirical observations on word coverage. Here, we observe that the larger the n-best list the higher proportion of words covered (*% covered*). Interestingly, we also observe similar trends for *% exact match*, but only if all n-best candidates are considered. It also interesting to note the difference in the impressive increase in *% exact match* from 1-best to all-best for NMT, which does not happen for SMT. These results show that for NMT larger beam sizes lead to more information content in translation candidates. Therefore, clever techniques to explore the space of hypotheses should lead to better translations.

Even though the NMT vs SMT figures are not directly comparable since the NMT and SMT systems are trained on different data, we note that despite the SMT system using a beam size of 200 and producing 500-best translation hypotheses, its translations have much lower word overlap than those from the NMT system with a beam size of 10 for English↔Czech. These results further corroborate the reasons for the insignificant gains obtained in the WMT16 SMT system Tuning shared task (Jawaid et al., 2016). In fact, if larger hypotheses spaces do not lead to more words that can potentially lead to translations that match the reference, the tuning algorithms do not have much to learn from.

4.2 Oracle Exploration

Based on the encouraging observations in the previous experiment with word overlap between candidates in the n-best list and the reference translation, here we attempt to quantify the potential gain from optimally exploring the space of hypotheses. We perform experiments assuming that we have an 'oracle' which helps us choose the best possible translation, under an evaluation metric against the reference, given an n-best list of translation hypotheses. This provides an upper-bound on the performance of the MT system. Positive results in this experiment will indicate that the MT system is capable of producing better translation candidates, but fails at scoring them as the best ones.

In this oracle experiment, the translation of a source sentence is chosen based on comparisons among the translation hypotheses and the reference translation – the oracle – under a

NEURAL MT	10-best		100-best		500-best	
	1-best	*all*	*1-best*	*all*	*1-best*	*all*
English→German						
%covered	53.99	62.75	53.99	71.93	53.83	77.69
% exact match	2.20	6.47	2.20	12.07	2.20	18.24
German→English						
%covered	57.32	65.98	57.43	74.42	57.43	79.55
% exact match	2.70	7.70	2.70	15.40	2.70	22.94
English→Czech						
%covered	45.97	55.27	45.85	65.61	45.72	72.55
% exact match	1.63	4.90	1.63	9.40	1.63	14.77
Czech→English						
%covered	52.30	61.26	52.33	70.24	51.92	75.61
% exact match	1.67	14.44	1.67	11.47	1.60	16.97

STATISTICAL MT	10-best		100-best		500-best	
(beam=200, distinct)	*1-best*	*all*	*1-best*	*all*	*1-best*	*all*
English→Czech						
% covered	39.20	46.58	39.20	54.05	39.20	57.86
% exact match	0.07	0.07	0.07	0.37	0.07	0.37
Czech→English						
% covered	48.35	54.79	48.35	60.30	48.35	62.89
% exact match	0.16	0.50	0.16	0.83	0.16	0.83

Table 2: Proportion of words overlapping between candidates and reference translations for different values of the n-best, as well as proportion of MT output sentences that exactly match the reference, considering either the 1-best or all the MT candidates in the n-best list.

certain MT evaluation metric. We consider the outputs of NMT systems for beam sizes of 10, 100 and 500 and with the following metrics: BLEU with n-gram max length $= 4$ and default brevity penalty settings, BEER2.0 with default settings, and ChrF with n-gram max length $= 6$ and $\beta = 3$. By exploring multiple metrics we will gain insights on how well different metrics do at spotting the best candidates: ideally, better metrics should lead to larger improvements from the original top translation.

Observations and Discussion We report the results of the oracle experiment in Figure 1. For each system, we report the relative improvement (delta) between the oracle translation chosen by the three metrics – BLEU, BEER and ChrF3 – compared to the 1-best of the system for a given n-best list size. Using any of the metrics we are able to find an alternative MT candidate which is better than the original 1-best translation, resulting in an overall increase in translation quality in all datasets. Larger improvements are obtained with larger beam sizes. However, while a large gain (almost double) is obtained from beam size 10 to 100, the rate of increase in improvement seems to drop from beam size 100 to 500, indicating that more additional translations are probably mostly spurious. This is consistent with the information content experiment in Section 4.1.

Kumar and Byrne (2004) reports that their MBR decoder leads to improvements only according to an evaluation metric that is also used as basis for their loss function. In our experiments, to better understand the relationship between the re-ranking metric and the final evaluation results, we further explore the oracle experiment by reporting results on the 500-best output for NMT, which brings the best gains in Figure 1, but focus on the proportion of improvement of the oracle translation over 1-best *across metrics*. In other words, we oracle re-rank using each given metric and evaluate the final 1-best translation set performance using all

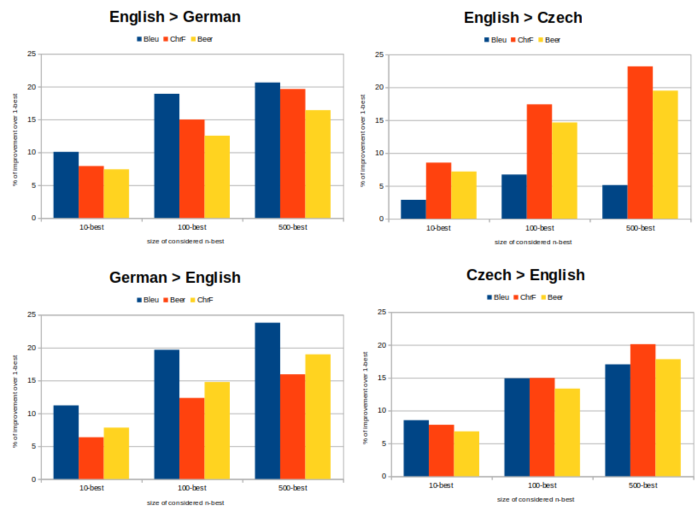

Figure 1: Proportion of improvement in NMT results according to MT evaluation metrics based on the oracle results over the original 1-best when the size of the beam is increased for decoding.

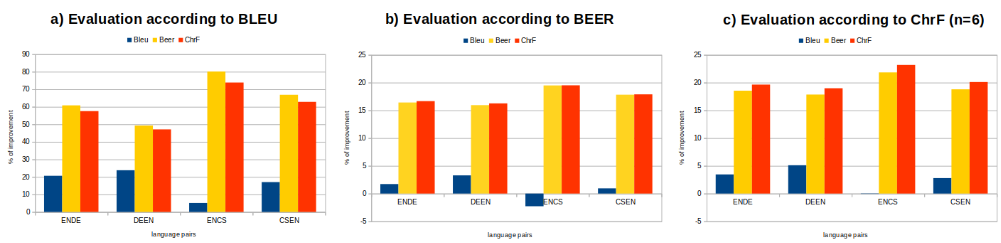

Figure 2: Focusing on the 500-best output for NMT, which brings the best gains in Figure 1, proportion of improvement of the oracle translation over the original 1-best when using different metrics for the oracle computation: ChrF3, BEER and BLEU. Re-ranking is done with one metric at a time, and the final performance is also measured with each of three metrics.

three metrics. This helps us assess the potential of each metric in selecting the best candidate. Figure 2 shows the results. Contrary to what was suggested in Kumar and Byrne (2004) for SMT, in chart (a) we see that the relative improvement is bigger in terms of the BLEU metric when using either BEER or ChrF3 to obtain the 1-best translation than using BLEU itself. We also observe in charts (b) and (c) that the character-based metrics always outperform BLEU and extract better 1-best translations. BLEU also seems to fail at identifying better MT candidates when translating into Czech, which is a morphologically rich language, while BEER and ChrF3 perform better. We note however that Kumar and Byrne (2004) also tune the log-linear loss function, while in our case we are just selecting the candidates directly based on a metric.

Since sentence length is a often problem in NMT, we measure the impact of using different evaluation metrics for oracle re-ranking on the sentence length of the 1-best translations chosen. In Figure 3 we report variation in terms of sentence length average for all NMT systems after the oracle translation selection with all three metrics, compared to the original 1-best translation for each setting. We notice that the average length of oracle BLEU translations does not seem to vary, however, an opposite trend is seen with BEER and ChrF3, which seem to make sentences shorter except for German→English. This is particularly interesting since i) we observe in Table 2 a better coverage with bigger beam size, and ii) we observe an overall large BLEU improvement our oracle experiments (Figure 2 (a)). This suggests that we are able to select translation candidates that might be shorter than the original 1-best, but most similar to the reference translation.

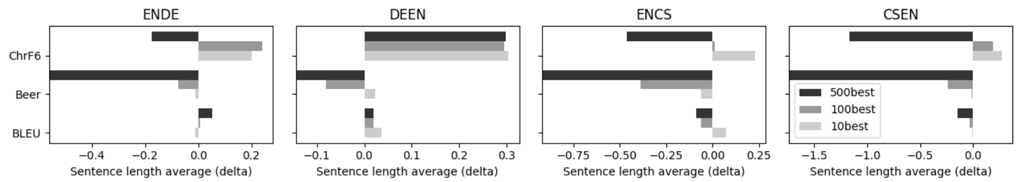

Figure 3: Delta in average sentence length for all NMT systems after 1-best oracle translation selection by each metric, compared to the average sentence length of the original 1-best.

5 Consensus-based n-best re-ranking

As was shown in the previous section, increasing the size of the beam generally leads to better word coverage and, more important, to higher chances of generating better translations among the resulting n-best lists. In what follows we propose an approach to automatically re-rank n-best lists to obtain better translations (without oracle translations).

Our approach is motivated by the work of DeNero et al. (2009) for SMT, where consensus-based MBR decoding is used to guide the choices of the decoder towards hypotheses that share partial translations. DeNero et al. (2009) experiment with different evaluation metrics (including BLEU) to measure similarity among hypotheses within a n-best list. We propose to empirically evaluate the contribution of consensus information in hypotheses in n-best lists from NMT systems. This is simpler than using consensual information at decoding time, but we believe that positive results at re-ranking stage will provide insights on whether or not this is a promising path to follow in NMT decoding.

Given an n-best list and a certain similarity metric, we compute the metric scores for each translation hypothesis against each of all $n-1$ other hypotheses in the n-best list. We then average the similarity scores of all $n-1$ translation hypotheses to obtain a single score for each translation hypothesis. We repeat this for all translation hypotheses and then sort the n-best list based on these scores, such that the top (best) translation will be one that is similar to more of the alternative candidates. Given that NMT systems produce translations are are "more likely" given the model, this essentially corresponds to selecting as best translation the one that is the most similar to all of $n-1$ the most likely translations. The size of the n-best list here is critical: the more hypotheses in the list, the less confident the NMT system will be on the bottom part of the list (less likely translations). However, longer n-best lists may provide stronger evidence for consensual analysis. This is a classical exploration-exploitation issue.

Another remark is that larger search spaces require much more time to compute the consensus-based re-ranking. We experiment with BLEU, BEER and ChrF3 as similarity metrics, since these are easily available and are either extremely popular (BLEU) or have proved to correlate well with human judgements on translation quality (in terms of similarity with a reference translation) in recent evaluation campaigns (BEER and ChrF3) (Bojar et al., 2016c). While each pair of translation hypotheses can be scored independently, which allows parallel processing, the running time for each metric to re-rank a complete n-best list is $O(n^2 \cdot k)$, where k is the size of the corpus and n the size of the n-best list. This may be very time consuming:

from hours up to a day[5] for easy-to-compute metrics such as BLEU or ChrF, to many days for more complex metrics such as BEER.

Automatic evaluation We start by evaluating our consensus-based re-ranking approach using BLEU as automatic evaluation metric. The results are shown in Table 3. A similar trend was observed using BEER and ChrF3 as similarity metrics, however we omit these results due to space constraints. Comparing the figures in this table against those in Table 1, we see that – under the same beam size – re-ranking seems to degrade the results in all cases with BLEU and ChrF, but not with BEER. An increase in BLEU scores can be observed for BEER-based re-ranking as longer beam sizes superior to 10 are used for the two language pairs where re-ranking under this metric was computed. It is not surprising to see that this improvement is only observed for BEER as similarity metric, even though the final evaluation is in terms of BLEU. This suggests that exploring other similarity metrics for the consensus analysis could be beneficial. Overall, re-ranking using BEER as similarity metric leads to the best results.

	English→German				**German→English**			
		re-ranked with				*re-ranked with*		
n-best	*baseline*	BLEU	BEER	ChrF3	*baseline*	BLEU	BEER	ChrF3
n=10	26.93	26.51	26.77	26.38	32.58	32.10	32.29	31.79
n=100	26.82	26.02	26.87	26.18	32.68	31.90	32.78	31.67
n=500	26.18	24.80	-	25.93	32.70	31.41	32.85	32.25
	English→Czech				**Czech→English**			
		re-ranked with				*re-ranked with*		
n-best	*baseline*	BLEU	BEER	ChrF3	*baseline*	BLEU	BEER	ChrF3
n=10	18.50	17.98	18.24	17.60	26.26	25.81	26.10	25.52
n=100	18.31	17.58	18.61	17.57	26.17	25.47	26.42	25.16
n=500	17.81	16.39	-	17.38	24.19	24.44	26.57	24.80

Table 3: BLEU scores of our consensus-based re-ranking strategy on the WMT16 test sets with NMT using n-best lists of sizes 10, 100 and 500. The scores are computed on the newly ranked 1-best NMT candidate against the reference translation. The *baseline* scores correspond to the original 1-best assessed towards the reference translation (see Table 1). The current implementation of BEER makes our consensus-based re-ranking extremely time consuming and virtually unfeasible, therefore we only show results for a subset of language pairs.

In Table 4 we illustrate some examples from the re-ranking approach. We observed that the consensus-based re-ranking produced interesting sentences that included syntactic re-orderings, new words, morphological variations and other nuances which were not captured by BLEU. This motivated us to perform human evaluation of the translations to more quantitatively compare the original 1-best versus the re-ranked 1-best.

Human evaluation We conducted a human evaluation using Appraise (Federmann, 2012), an open-source web application for manual evaluation of MT output. Appraise collects human judgements on translation output, implementing annotation tasks such as quality checking, error classification, manual post-editing and, in our case, translation ranking. For a list of up to four systems' outputs for each source sentence, we requested human annotators to rank the set of MT candidates from the best to the worst, allowing for ties, based on both the source sentence and reference translation. If two system outputs are the same, the MT candidate was displayed once and the same rank was assigned to both systems.

For this evaluation, we selected a subset of our systems based on our automatic evaluation results: for each metric used for re-ranking in each language pair, we chose the systems that

[5]Indicative time it took to re-rank a corpus of 3,000 sentences, with $n = 500$ on a 40-cores CPU server.

German→English	
SRC:	Das rund zehn bis zwölf Millionen Euro teure Vorhaben steht seit Monaten in der Diskussion.
REF:	The € 10 - 12 million project has been under discussion for months.
Baseline:	the EUR 10 million project has been under discussion for months.
BEER:	the approximately EUR 10 to 12 million projects has been under discussion for months
ChrF3:	the EUR 10 million euro project has been under discussion for several months.
BLEU:	the projects around ten to twelve million euros have been discussed for months.
Czech→English	
SRC:	Navíc jsem si ze života odnesl zkušenost, že zasahování do ekosystému nevede k úspěchu a jednoho škůdce může nahradit druhý.
REF:	Furthermore, in my experience, interfering with the ecosystem does not lead to success and one pest can replace another.
Baseline:	moreover, I have learned from life that interfering with an ecosystem does not lead to success, and one pest can replace another.
BEER:	moreover, I have learned from my life that it is not possible to succeed in an ecosystem, and one can replace one of the pests.
ChrF3:	moreover, I have learned from life that interfering with an ecosystem does not lead to success, and one pest can replace one another.
BLEU:	moreover, I have learned from life that interfering with an ecosystem does not lead to success, and one pest can replace one.

Table 4: Examples of alternative MT candidates chosen by consensus from n-best lists (with $n = 500$). Boxes highlight the main differences between the reference translation, the baseline (i.e. the original 1-best) and an alternative translation chose by our consensus re-ranking approach using BLEU, BEER or ChrF.

performed the best according to the three metrics (averaged ranking among the three), along with the original 1-best.

Each human translator was asked to complete at least one hit of twenty annotation tasks. Incomplete hits were discarded from the evaluation. We collected 3,016 complete ranking results over the four NMT systems (159 for English→Czech, 1,365 for Czech→English, 911 for English→German, 581 for German→English), from 208 annotators.

We borrowed a method from the WMT translation shared task to generate a global ranking of systems from these judgements. Table 5 reports the ranking results according to the Expected Wins method[6] for the four language pairs. The first column ($\#_m$) indicates the ranking of the systems amongst themselves according to the three automatic metrics, while the third column (range) indicates the ranking from the human evaluation. For example, for English→German, the *BLEU-100best* system was ranked first amongst the four by all three metrics, but it was ranked last by human annotators.

[6] https://github.com/keisks/wmt-trueskill/blob/master/src/infer_EW.py

English→German

$\#_m$	score	range	system
4	0.578	1-2	BEER-100best
2	0.529	1-3	Baseline (10best)
3	0.505	2-3	ChrF3-10best
1	0.388	4	BLEU-100best

German→English

$\#_m$	score	range	system
4	0.559	1-3	BEER-500best
2	0.546	1-3	Baseline (10best)
3	0.525	1-3	ChrF3-10best
1	0.393	4	BLEU-500best

English→Czech

$\#_m$	score	range	system
2	0.583	1-3	BEER-100best
4	0.532	1-3	ChrF3-100best
1	0.493	1-4	BLEU-100best
3	0.372	3-4	Baseline (100best)

Czech→English

$\#_m$	score	range	system
4	0.526	1-3	BEER-10best
3	0.522	1-2	ChrF3-500best
2	0.508	1-3	Baseline (500best)
1	0.453	3-4	BLEU-500best

Table 5: Results of the human evaluation for NMT. Systems are sorted according to human assessments while $\#_m$ indicates the overall ranking of a system according to all three automatic metrics. Scores and ranges are obtained with the Expected Wins method (Sakaguchi et al., 2014). Lines between systems indicate clusters. Systems within a cluster are considered tied. In gray are systems which have not significantly outperformed the baseline.

Our first observation is that the consensus-based re-ranking with BEER outperforms the other two metrics for all the language pairs, confirming the results of the automatic evaluation. Except for Czech→English, systems always benefit from a beam size larger than 10, which suggests that we should consider exploiting a larger search spaces in NMT. Another interesting outcome of the human evaluation is the ranking of our systems, which for most of the language pairs refutes the ranking according to the automatic evaluation. Although those metrics are known to be well correlated with human judgements, it seems that humans have different perceptions on the quality of the translations.

6 Conclusions

In this paper we reported our experiments and results on the influence of the beam size in NMT. While traditional approaches in NMT rely on smaller beam sizes or use greedy implementations, our paper strongly motivates using a larger beam size. We investigate the informativeness of larger beam size and highlighted the potential to improve translation quality by exploring larger hypotheses spaces using an oracle experiment. Motivated by substantial potential gains in both informativeness and oracle-based hypotheses re-ranking, we proposed a consensus-based NMT n-best re-ranking approach, with insights into the use of different metrics to capture consensus-based information. Our contribution strongly suggests further work in NMT to explore larger beams and n-best lists.

Acknowledgements

This work was supported by the QT21 project (H2020 No. 645452).

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Blackwood, G., de Gispert, A., and Byrne, W. (2010). Efficient path counting transducers for minimum bayes-risk decoding of statistical machine translation lattices. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 27–32, Uppsala, Sweden.

Blatz, J., Fitzgerald, E., Foster, G., Simona Gandrabur, C. G., Kulesza, A., Sanchis, A., and Ueffing, N. (2003). Confidence estimation for machine translation. Technical report, Johns Hopkins University.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016a). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Bojar, O., Dušek, O., Kocmi, T., Libovický, J., Novák, M., Popel, M., Sudarikov, R., and Variš, D. (2016b). CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *19th International Text, Speech and Dialogue Conference*, Brno, Czech Republic. Springer Verlag.

Bojar, O., Graham, Y., Kamran, A., and Stanojevic, M. (2016c). Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991.

DeNero, J., Chiang, D., and Knight, K. (2009). Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Conference on Natural Language Processing of the AFNLP*, pages 567–575.

Duh, K. and Kirchhoff, K. (2008). Beyond log-linear models: boosted minimum error rate training for n-best re-ranking. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 37–40.

Duh, K., Sudoh, K., Tsukada, H., Isozaki, H., and Nagata, M. (2010). N-best reranking by multitask learning. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 375–383, Uppsala, Sweden. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2.

Ehling, N., Zens, R., and Ney, H. (2007). Minimum Bayes risk decoding for BLEU. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 101–104, Prague, Czech Republic.

Federmann, C. (2012). Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Gimpel, K., Batra, D., Dyer, C., and Shakhnarovich, G. (2013). A systematic exploration of diversity in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hasan, S., Zens, R., and Ney, H. (2007). Are very large n-best lists useful for smt? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 57–60, Rochester, New York.

Jawaid, B., Kamran, A., Stanojević, M., and Bojar, O. (2016). Results of the wmt16 tuning shared task. In *Proceedings of the First Conference on Machine Translation*, pages 232–238, Berlin, Germany.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Demon Session*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton, Canada.

Kumar, S. and Byrne, W. (2004). Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*, pages 169–176, Boston, MA.

Lala, C., Madhyastha, P., Wang, J., and Specia, L. (2017). Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. *The Prague Bulletin of Mathematical Linguistics*.

Lambert, P. and Banchs, R. (2006). Tuning machine translation parameters with spsa. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 190–196.

Li, J. and Jurafsky, D. (2016). Mutual information and diverse decoding improve neural machine translation. *CoRR*, abs/1601.00372.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(3):8:1–8:49.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Neubig, G., Morishita, M., and Nakamura, S. (2015). Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. *CoRR*, abs/1510.05203.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A. M., Kumar, S., Shen, L., Smith, D., Eng, K., et al. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*, pages 161–168.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Popovic, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 392–395.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11.

Sennrich, R., Haddow, B., and Birch, A. (2015a). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2015b). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany.

Shen, L., Sarkar, A., and Och, F. J. (2004). Discriminative reranking for machine translation. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*, pages 177–184.

Shu, R. and Nakayama, H. (2017). Later-stage Minimum Bayes-Risk Decoding for Neural Machine Translation. *ArXiv e-prints*.

Sokolov, A., Wisniewski, G., and Yvon, F. (2012a). Computing lattice bleu oracle scores for machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 120–129, Avignon, France.

Sokolov, A., Wisniewski, G., and Yvon, F. (2012b). Non-linear n-best reranking with few features. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas*, pages 1–10, San Diego, CA.

Specia, L., Sankaran, B., and Graças Volpe Nunes, M. (2008). n-best reranking for the efficient integration of word sense disambiguation and statistical machine translation. *Lecture Notes in Computer Science*, 4919:399–410.

Stahlberg, F., de Gispert, A., Hasler, E., and Byrne, B. (2017). Neural machine translation by minimising the bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 362–368, Valencia, Spain.

Stahlberg, F., Hasler, E., Waite, A., and Byrne, B. (2016). Syntactically guided neural machine translation. *arXiv preprint arXiv:1605.04569*.

Stanojevic, M. and Sima'an, K. (2014). Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

Tromble, R. W., Kumar, S., Och, F., and Macherey, W. (2008). Lattice minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Wisniewski, G., Allauzen, A., and Yvon, F. (2010). Assessing phrase-based translation models with oracle decoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 933–943, Cambridge, Massachusetts.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, Y., Hildebrand, A. S., and Vogel, S. (2006). Distributed language modeling for n-best list re-ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Sydney, Australia.

Confidence through Attention

Matīss Rikters matiss@lielakeda.lv
Faculty of Computing, University of Latvia

Mark Fishel fishel@ut.ee
Institute of Computer Science, University of Tartu, Estonia

Abstract

Attention distributions of the generated translations are a useful bi-product of attention-based recurrent neural network translation models and can be treated as soft alignments between the input and output tokens. In this work, we use attention distributions as a confidence metric for output translations. We present two strategies of using the attention distributions: filtering out bad translations from a large back-translated corpus, and selecting the best translation in a hybrid setup of two different translation systems. While manual evaluation indicated only a weak correlation between our confidence score and human judgments, the use-cases showed improvements of up to 2.22 BLEU points for filtering and 0.99 points for hybrid translation, tested on English↔German and English↔Latvian translation.

1 Introduction

Neural machine translation (NMT) has recently redefined the state-of-the-art in machine translation (Sennrich et al., 2016a; Wu et al., 2016a), with one of the ground-breaking innovations that enabled this being the introduction of the attention mechanism (Bahdanau et al., 2014). It enables the model to find parts of a source sentence that are relevant to predicting a target word (pay attention), without the need to form these parts as a hard segment explicitly. Decoding sentences with the attention-based model resulted in a useful by-product – soft alignments between tokens of source and target sentences. These can be used for many purposes, such as replacing unknown words with back-off translations from a dictionary (Jean et al., 2015) and visualizing the soft alignments (Rikters et al., 2017).

In this paper, we propose using the attention alignments as an indicator of the translation output quality and the confidence of the decoder. We define metrics of confidence that detect and penalize under-translation and over-translation (Tu et al., 2016) as well as input and output tokens with no clear alignment, assuming that all these cases most likely mean that the quality of the translation output is bad.

We apply these attention-based metrics to two use-cases: scoring translations of an NMT system and filtering out the seemingly unsuccessful ones, and comparing translations from two different NMT systems, in order to select the best one.

The structure of this paper is as follows: Section 2 summarizes related work in back-translating with NMT, machine translation combination approaches and confidence estimation. Section 3 introduces the problem of faulty attention distributions and a way to quantify it as a confidence score. Sections 4 and 5 outline the two use-cases for this score – translation filtering and hybrid selections. Finally, we conclude in Section 6 and mention directions for future work in Section 7.

2 Related Work

Back-translation of Monolingual Data

One of the first uses of back-translation of monolingual data as an additional source of training data was reported by (Sennrich et al., 2016a) in their submission for the WMT16 news translation shared task. They translated target-language monolingual corpora into the source language of the respective language pair, and then used the resulting synthetic parallel corpus as additional training data. They performed experiments in ranges from 2 million to 10 million back-translated sentences and reported an increase of 2.2 - 7.7 BLEU (Papineni et al., 2002) for translating between English and Czech, German, Romanian and Russian. The authors also experimented with different amounts of back-translated data and found that adding more data gradually improves performance.

In a later paper Sennrich et al. (2016b) explored other methods of using monolingual data. They experimented with adding an enormous amount of monolingual sentences as targets without any sources to the parallel corpus and compared that to performing back-translation on a part of the monolingual data. While both methods outperform using just parallel data, the back-translated synthetic parallel corpus is a much more powerful addition than the mono data alone.

Pinnis et al. (2017) experimented with using large and even larger amounts of back-translated data and came to a conclusion that any amount is an improvement, but using double the amount gives lower results, while still better than not using any at all. These results hint that it may be possible to get even better results when using only the part of the data selected with some criterion. One of the aims of our work is to provide one such criterion.

Machine Translation System Combination

Zhou et al. (2017) used attention to combine outputs from NMT and SMT systems. The authors first trained intermediate NMT, SMT and hierarchical SMT systems with one-half of the training data. Afterwards, they used each system to translate the target side of the other half of the training data. Finally, the three translated parts as source sentence variants along side the clean target sentence were used for training the combination neural network. This approach gave the network more choices of where to pay attention and which parts should be ignored in the training process. They perform experiments on Chinese→English and report BLEU score improvement by 5.3 points over the best single system and 3.4 points over traditional MT combination methods.

Peter et al. (2016) perform MT system combination in a more traditional manner - using confusion networks. They use 12 different SMT and NMT systems to generate hypothesis translations, align and reorder each hypothesis to match one skeleton hypothesis, creating a confusion network. For the final output is generated by finding the best path in the network. The authors report an improvement of 1.0 BLEU compared to the best single system, translating from English into Romanian.

Translation Confidence Metrics

Lately the idea of modeling coverage in NMT was introduced, for example, Tu et al. (2016) integrate it directly into the attention mechanism and report improved translation quality as a result. On the simpler side of things, Wu et al. (2016b) perform tests with a baseline attention that uses an additional coverage penalty at decoding time; they report no improvement compared to the common length normalization. Our metrics are partially motivated by the coverage penalty, though we apply them at the post-translation stage to determine the confidence of the decoder and the quality of the already made translation, which makes it applicable regardless of which software or approach were used.

Another closely related task is quality estimation. The dominating approach there is collecting post-edits and training a machine learning model to predict the quality score or classify translations into usable/not, near-perfect/not, etc (Bach et al., 2011; Felice and Specia, 2012). The main similarity between our work and quality estimation is their usage of glass-box features (i.e. information about the MT system or the decoder's internal parameters). While our approach does not cover all aspects of quality estimation, it requires no data or training and can be applied to any language and neural machine translation system.

3 Penalizing Attention Disorders

Before describing the confidence metrics based on attention weights, here is a brief overview of the NMT architecture where the attention weights come from.

3.1 Source of Attention

Our work is built around the encoder-decoder machine translation approach (Sutskever et al., 2014; Cho et al., 2014) with an attention mechanism (Bahdanau et al., 2014). In this approach the source tokens are learned to be represented by an encoder, which consists of an embedding layer and a bi-directional LSTM or GRU layer (or 8, Wu et al., 2016b), the outputs of which serve as the learned representation.

There is also a decoder that consists of another layer (or 8, ibid.) of LSTM/GRU cells, with an output layer for predicting the softmax-encoded raw probability distribution of each output word, one at a time. The state of the decoder layer(s) and thus the output distribution depends on the previous recurrent states, the previously produced output word and a weighted sum of the representations of the source sentence tokens. The weights in this sum are generated for every output word by the attention mechanism, which is a feed-forward neural network with the previous state of the decoder and each input word representation as input and the raw weight of that word for the next state as output. Finally, the attention weights are normalized as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}$$

where e_{ij} is the raw predicted weight and α_{ij} – the final attention weight between the input token j and output token i.

Once the encoder-decoder network has been trained, it can be used to produce translations by predicting the probability for each next word, which can serve as the basis for sampling, greedy search or beam search (Sennrich et al., 2017). We refer the reader for a complete description to (Bahdanau et al., 2014) and ourselves turn on to the main topic of the paper that uses the weights α_{ij} to estimate the confidence of the translations.

Together with the translation, it is also possible to save the attention values between the input tokens and each produced output token. These values can be interpreted as the influence of the input token on the output token, or the strength of the connection between them. Thus, weak or dispersed connections should intuitively indicate a translation with low confidence, while high values and strong connections between one or two tokens on both sides should indicate higher confidence. Next, we present our take at formalizing this intuition.

3.2 Measuring Attention

Figure 1 shows an example of a translation that has little or nothing to do with the input, a frequent occurrence in NMT. Besides the text of the translation, it is clear already by looking at the attention weights of this pair that the translation is weak:

- some input tokens (like the sentence-final full-stop) are most strongly connected to several unrelated output tokens, in other words, their coverage is too high,

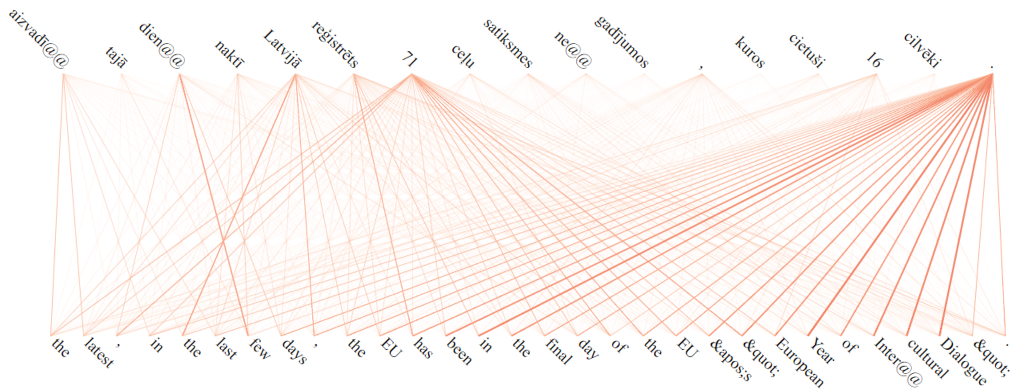

Figure 1: Attention alignment visualization of a bad translation. Reference translation: *71 traffic accidents in which 16 persons were injured have happened in Latvia during the last 24 hours.*, hypothesis translation: *the latest , in the last few days , the EU has been in the final day of the EU 's " European Year of Intercultural Dialogue ".* $CDP = -0.900$, $AP_{out} = -2.809$, $AP_{in} = -2.137$, $Total = -5.846$.

- most of the input token attentions, as well as some output token attentions, are highly dispersed, without one or two clear associations on the counterpart.

On the other hand, a picture like Figure 2 intuitively corresponds to a good translation, with strongly focused alignments. It is this intuition that our metrics formalize: penalizing translations with tokens with a total coverage of not just below but much higher than 1.0, as well as tokens with a dispersed attention distribution.

Coverage Deviation Penalty

Previous work (Wu et al., 2016b) defines a coverage penalty, which is meant to punish translations for not paying enough attention to input tokens:

$$CP = \beta \sum_j \log(\min(\sum_i \alpha_{ji}, 1.0)),$$

where i is the output token index, j – the input token index, β is used to control the influence of the metric and CP – the coverage penalty.

The first part of our metric draws inspiration from the coverage penalty; however, it penalizes not just lacking attention but also too much attention per input token. The aim is to penalize the sum of attentions per input token for going too far from 1.0[1], so tokens with total attention of 1.0 should get a score of 0.0 on the logarithmic scale, while tokens with less attention (like 0.2) or more attention (like 2.5) should get lower values. We thus define the coverage deviation penalty:

$$CDP = -\frac{1}{J} \sum_j \log\left(1 + (1 - \sum_i \alpha_{ji})^2\right),$$

where J is the length of the input sentence. The metric is on a logarithmic scale, and it is normalized by the length of the input sentence in order to avoid assigning higher scores to shorter sentences[2]. See examples of the CDP metric's values on Figures 1 and 2.

[1]This could be replaced with the token's expected fertility, which we leave for future work

[2]This is not required for choosing translations of the same sentence by the same system, but is required in our

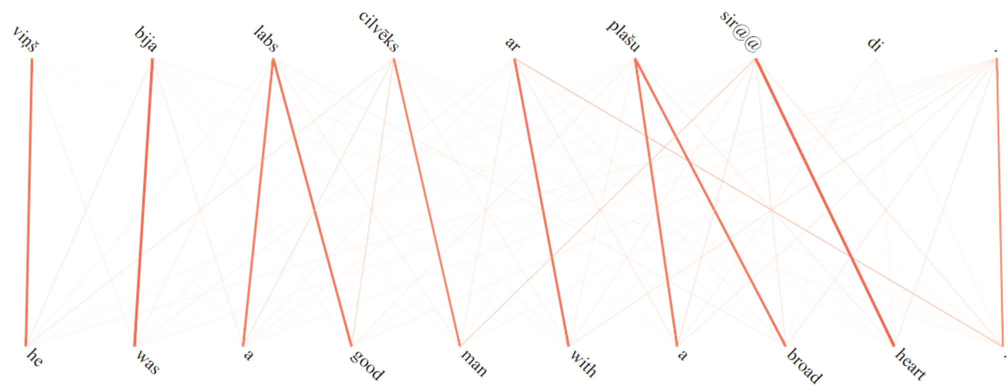

Figure 2: Attention alignment visualization of a good translation. Reference translation: *He was a kind spirit with a big heart.*, hypothesis translation: *he was a good man with a broad heart*. $CDP = -0.099$, $AP_{out} = -1.077$, $AP_{in} = -0.847$, $Total = -2.024$.

Absentmindedness Penalty

However, it is not enough to simply cover the input, we conjecture that more confident output tokens will allocate most of their attention probability mass to one or a small number of input tokens. Thus the second part of our metric is called the absentmindedness penalty and targets scattered attention per output token, where the dispersion is evaluated via the entropy of the predicted attention distribution. Again, we want the penalty value to be 1.0 for the lowest entropy and head towards 0.0 for higher entropies.

$$AP_{out} = -\frac{1}{I} \sum_i \sum_j \alpha_{ji} \cdot \log \alpha_{ji}$$

The values are again on the log-scale and normalized by the source sentence length I.

The absentmindedness penalty can also be applied to the input tokens after normalizing the distribution of attention per input token, resulting in the counter-part metric AP_{in}. This is based on the assumption that it is not enough to cover the input token, but rather the input token should be used to produce a small number of outputs. See examples of both metric's values on Figures 1 and 2.

Finally, we combine the coverage deviation penalty with both the input and output absentmindedness penalties into a joint metric via summation:

$$confidence = CDP + AP_{out} + AP_{in}$$

Next, we evaluate the metrics directly against human judgments and indirectly by applying them to filtering translations and plugging them into a sentence-level hybrid translation scheme.

3.3 Human Evaluation

It is clear that the defined metrics only paint a partial picture, since they rely on the attention weights only. For instance, they do not evaluate the lexical correspondence between the source and hypothesis, and more generally, being confident does not mean being right. We wanted to find out how much confidence in our case correlates with translation quality.

experiments described in the next sections.

To do so we asked human volunteers to perform pairwise ranking of translations from two baseline NMT systems: one done with Nematus (Sennrich et al., 2017) and the other – with Neural Monkey (Helcl and Libovickỳ, 2017). The translations and measurements were done for English-Latvian and Latvian-English, using corpora from the news translation shared task of WMT'2017; further details can be found in Section 4. We selected 200 random sentences for both translation directions and these were given to native Latvian speakers for evaluation. The MT-EQuAl (Girardi et al., 2014) tool was used for the evaluation task. The evaluators were shown one source sentence at a time along with the two different translations. They were instructed to assign one of five categories for each translation: "worst", "bad", "ok", "good" or "best", noting that both may be categorized as equally "good" or "bad", etc. Differing judgments for the same sentence were averaged. All 200 sentences were annotated by at least one human annotator.

It makes more sense to treat the results as relative comparisons, not absolute scores, as the annotators only see two translations at a time. We use these comparisons to compute the Kendall rank correlation coefficient (Kendall, 1938) by only looking at the pairs where human scores differ. Since we only have comparisons for each pair and not between different sentences, the coefficient is computed as

$$\tau = \frac{pos - neg}{pos + neg},$$

where pos is the number of pairs where the metric agrees with the human judgment and neg is the number of pairs where they disagree.

The results are presented in Table 1, and as we can see they indicate weak correlation, with the absolute values of τ between 0.012 and 0.200.

Language pair	CDP	AP$_{in}$	AP$_{out}$	Overall
En→Lv	0.099	0.074	0.123	0.086
Lv→En	-0.012	-0.153	-0.200	-0.153

Table 1: The Kendall's Tau correlation between human judgments and the confidence scores.

Let us look closer at where the metrics disagree with human judgments. Figure 3 shows an example of a translation which was rated highly by human annotators but poorly with our metrics. While the sentence is a good translation, it does not follow the source word-by-word. Some subword units and functional words do not have a clear alignment, even though they are understood/generated correctly. This means that one problem with our metrics is that they might be over-penalizing translations that deviate from a direct literal translation.

Next, we continue with the experiments of using our metrics to filter synthetic data and to select translations in a hybrid MT scenario.

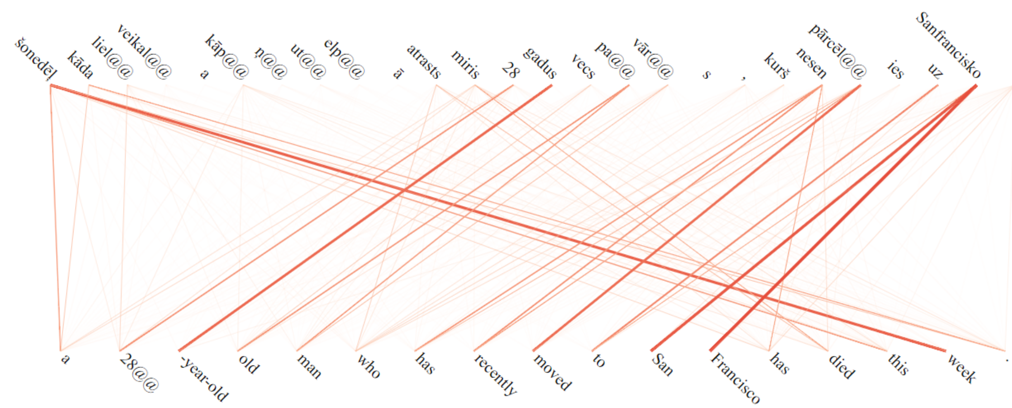

Figure 3: Attention alignment visualization of a bad translation. Reference translation: *a 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week .*, hypothesis translation: *a 28-year-old old man who has recently moved to San Francisco has died this week .*, $CDP = -0.250$, $AP_{out} = -1.740$, $AP_{in} = -1.46$, $Total = -3.45$.

4 Filtering Back-translated Data

4.1 Baseline Systems and Data

Our baseline systems were trained with two NMT frameworks - Nematus (NT) (Sennrich et al., 2017) and Neural Monkey (NM) (Helcl and Libovickỳ, 2017). For all NMT models we used a shared subword unit vocabulary (Sennrich et al., 2016c) of 35000 tokens, clip the gradient norm to 1.0 (Pascanu et al., 2013), dropout of 0.2, trained the models with Adadelta (Zeiler, 2012) and performed early stopping after 7 days of training. For models with each NMT framework we used the default settings as mentioned in the frameworks documentation:

- For NT models we used a maximum sentence length of 50, word embeddings of size 512, and hidden layers of size 1000. For decoding with NT we used beam search with a beam size of 12.
- For NM models we used a maximum sentence length of 70, word embeddings and hidden layers of size 600. For decoding with NM a greedy decoder was used.

Training, development and test data for all systems in both language pairs and translation directions was used from the WMT17 news translation task [3]. For the baseline systems, we used all available parallel data, which is 5.8 million sentences for En↔De and 4.5 million sentences for En↔Lv.

4.2 Back-translating and Filtering

We used our baseline En→Lv and Lv→En NM and NT systems to translate all available Latvian monolingual news domain data - 6.3 million sentences in total from *News Crawl: articles from 2014, 2015, 2016*, and the first 6 million sentences from the English *News Crawl 2016*. Much more monolingual data was available from other domains aside from news. Since the development and test data was of the news domain, we only used that, considering it as in-domain data for our systems.

[3] EMNLP 2017 Second Conference on Machine Translation - http://www.statmt.org/wmt17/

For each translation, we used the attention provided from the NMT system to calculate our confidence score, sorted all translations according to the score and selected the top half of the translations along with the corresponding source sentences as the synthetic parallel corpus. We used only the full confidence score (combination of CDP, AP_{out} and AP_{in}) for filtering instead of each individual score due to its smoother overall correlation with human judgments. In between, we also removed any translation that contained any *<unk>* tokens.

To compare attention-based filtering with a different method, we trained a CharRNN[4] language model (LM) with 4 million news sentences from each of the target languages. We used these LMs to get perplexity scores for all translations, order them and get the *better half*. Table 2 summarizes how much human evaluation overlaps with each of the filtering methods. The final row indicates how much both filtering methods overlap with each other. While results from either approach don't look overly convincing, the LM-based approach has been proven to correlate with human judgments close to the BLEU score and is a good evaluation method for MT without reference translations (Gamon et al., 2005). Therefore the attention-based approach that does not require training of an additional model and overlaps with human judgments to approximately the same level should be more desirable.

Filtering Method	En→Lv	Lv→En
LM-based overlap with human	58%	56%
Attention-based overlap with human	52%	60%
LM-based overlap with Attention-based	34%	22%

Table 2: Human judgment overlap results on 200 random sentences from the *newsdev2017* dataset compared to filtering methods.

4.3 NMT with Filtered Synthetic Data

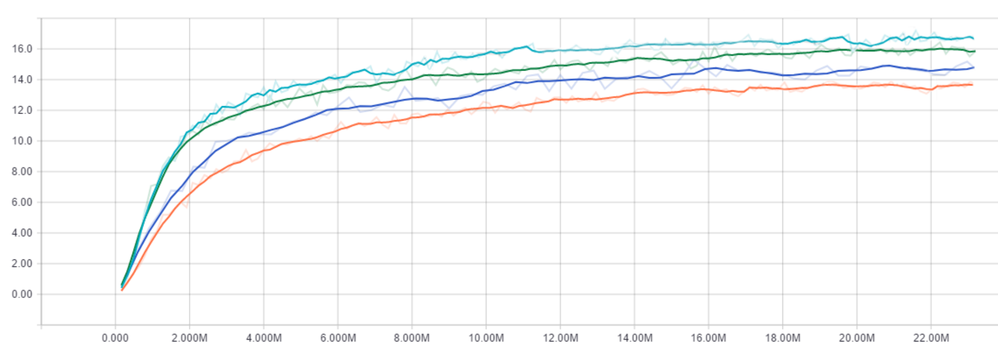

Figure 4: Automatic evaluation progression of Lv→En experiments on validation data. Orange – baseline; dark blue — with full back-translated data; green – with LM-filtered back-translated data; light blue – with attention-filtered back-translated data.

We shuffled each synthetic parallel corpus with the baseline parallel corpora and used them to train NMT systems. In addition to the baseline and two types of filtered BT synthetic data, we also trained a system with the full BT data for each translation direction. Figure 4

[4]Multi-layer Recurrent Neural Networks (LSTM, GRU, RNN) for character - level language models in Torch https://github.com/karpathy/char-rnn

shows a combined training progress chart for Lv→En on the full *newsdev2017* dataset that was used as the development set for training. Here the differences between all four approaches are clearly visible. Further results on a subset of *newsdev2017* and the full *newstest2017* dataset are summarized in Table 3. While for Lv→En and En↔De the attention-based approach is the clear leader, for En→Lv it falls behind the LM filtered version. We were not able to identify a clear reason for this and leave it for the future work. As expected, adding BT synthetic training data allows to get higher BLEU scores in all cases. It can be observed that filtering out half of the badly translated data and keeping only the best translations either does not decrease the final output quality in some cases or even further increase the quality in others, when using the LM. With filtering by attention, the results are more inconsistent - even higher in one direction while deterioration in the other. A reason for this could be that for Lv→En attention-based filtering the similarity with human judgments was higher than for En→Lv (Table 2), and it was also more different from the LM-based one. While for the other direction it is the other way around.

	BLEU							
Dataset	Dev	Test	Dev	Test	Dev	Test	Dev	Test
System	En→Lv		Lv→En		En→De		De→En	
Baseline	8.36	11.90	8.64	12.40	25.84	20.11	30.18	26.26
+ Full Synthetic	9.42	13.50	9.01	13.81	28.97	22.68	34.82	29.35
+ LM-Filtered Synthetic	9.75	13.52	9.45	14.30	29.59	23.48	34.47	29.42
+ Attn.-Filtered Synth.	8.99	12.76	**11.23**	**14.83**	**30.19**	23.16	**35.19**	**29.47**

Table 3: Experiment results in BLEU for translating between English↔Latvian with different types of back-translated data using development (200 random sentences from *newsdev2017*) and test (*newstest2017*) datasets.

5 Attention-based Hybrid Decisions

We translated the development set with both baseline systems for each language pair in each direction. The hybrid selection of the best translation was performed similarly to filtering, where we discarded the worst-scoring half of the translations. In the hybrid selection, we used the same score to compare both translations of a source sentence and choose the better one. Results of the hybrid selection experiments are summarized in Table 4. For translating between En↔Lv, where the difference between the baseline systems is not that high (0.06 and 1.55 BLEU), the hybrid method achieves some meaningful improvements. However, for En↔De, where differences between the baseline systems are bigger (3.46 and 4.46 BLEU), the hybrid drags both scores down.

	BLEU			
System	En→De	De→En	En→Lv	Lv→En
Neural Monkey	18.89	26.07	13.74	11.09
Nematus	22.35	30.53	13.80	12.64
Hybrid	20.19	27.06	**14.79**	**12.65**
Human	23.86	34.26	15.12	13.24

Table 4: Hybrid selection experiment results in BLEU on the development dataset (200 random sentences from *newsdev2017*).

The last row of the results Table 4 shows BLEU scores for the scenario when human annotator preferences were used to select each output sentence. An overview of human evaluator preferred translation selections is visible in Table 5. The results show that out of all translations the human evaluators deliberately prefer one or the other system. Aside from En→Lv, where a slight tendency towards Neural Monkey translations can be observed, all others look more or less equal. This highly contrasts with the BLEU scores from Table 4, where in both translation directions from English human evaluators prefer the lower-scoring system more often than the higher-scoring one. The final row of Table 5 shows how much our attention-based score matches the human judgments in selecting the best translation.

System	En→De	De→En	En→Lv	Lv→En
Neural Monkey	54%	42%	61.5%	47%
Nematus	46%	58%	38.5%	53%
Overlaps with hybrid selection	57%	47%	62.5%	51%

Table 5: Human evaluation results on 200 random sentences from the *newsdev2017* dataset compared to attention-hybrid selection.

6 Conclusions

In this paper, we described how attentional data from neural machine translation systems can be useful for more than just visualizations or replacing specific tokens in the output. We introduced an attention-based confidence score that can be used for judging NMT output. Two applications of using attentional data were investigated and compared to similar approaches. We used a smaller dataset to perform manual evaluation and compared that to all automatically obtained results. Our experiments showed interesting results and some increases in automated evaluation, as well as a good correlation with human judgments.

In addition to the methods described in this paper, we release open-source scripts[5] for (1) scoring, ordering and filtering NMT translations, (2) performing hybrid selections between two different NMT outputs of the same source, and (3) software for inspecting attention alignments that the NMT systems produce in the translation process (used for Figures 1 and 2). We also provide all development subsets that we used for manual evaluation with anonymized human annotations.

7 Future Work

This paper introduced the first steps in using NMT attention for less obvious intentions. It seemed that the attention score can complement the LM perplexity score in distinguishing good from bad translations. An idea for future experiments could be combining these scores to achieve a higher correlation with human judgments.

Additional improvements can be made to the hybrid decisions as well. Since the score represents the systems *confidence*, a badly trained NMT system can be more confident about a bad translation than a good system about a decent translation. While a hybrid combination of two similar quality NMT systems did put the attention score to good use, in the case with different quality systems the confidence of the weaker one was a pitfall. This indicates that the confidence score could be used in ensemble with a quality estimation score or used as a feature in training an MT quality estimation system.

For filtering synthetic back-translated data we dropped the worst-scoring 50% of the data, but this threshold may not be optimal for all scenarios. Several paths worth more exploration

[5]Confidence Through Attention - https://github.com/M4t1ss/ConfidenceThroughAttention

include exploring the effects of different static thresholds (e.g. 30% or 70%) or clustering the data by confidence score and dropping the lowest-scoring one or two clusters. Another path worth exploring for filtering would be to see how filtering by each individual score (CDP, AP_{in}, AP_{out}) compares to filtering by confidence.

In the near future, we also plan to supplement an attention inspection tool so that it displays confidence metrics and additional visualizations based on these scores.

References

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 211–219, Portland, Oregon, USA. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (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)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Felice, M. and Specia, L. (2012). Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Canada.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of EAMT*, pages 103–111.

Girardi, C., Bentivogli, L., Farajian, M. A., and Federico, M. (2014). MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123.

Helcl, J. and Libovickỳ, J. (2017). Neural monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107(1):5–17.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–89.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Peter, J.-T., Alkhouli, T., Ney, H., Huck, M., Braune, F., Fraser, A., Tamchyna, A., Bojar, O., Haddow, B., Sennrich, R., Blain, F., Specia, L., Niehues, J., Waibel, A., Allauzen, A., Aufrant, L., Burlot, F., Knyazeva, E., Lavergne, T., Yvon, F., Frank, S., Daiber, J., and Pinnis, M. (2016). The QT21/HimL Combined Machine Translation System. *Proceedings of the First Conference on Machine Translation (WMT 2016), Volume 2: Shared Task Papers*, 2:344—-355.

Pinnis, M., Krislauks, R., Deksne, D., and Miks, T. (2017). Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *International Conference on Text, Speech, and Dialogue*, pages 20–27. Springer.

Rikters, M., Fishel, M., and Bojar, O. (2017). Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(3):in print.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Neural Machine Translation of Rare Words with Subword Units. In *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, Montreal, Canada.

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016a). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016b). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhou, L., Hu, W., Zhang, J., and Zong, C. (2017). Neural System Combination for Machine Translation.