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 pre-specified 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

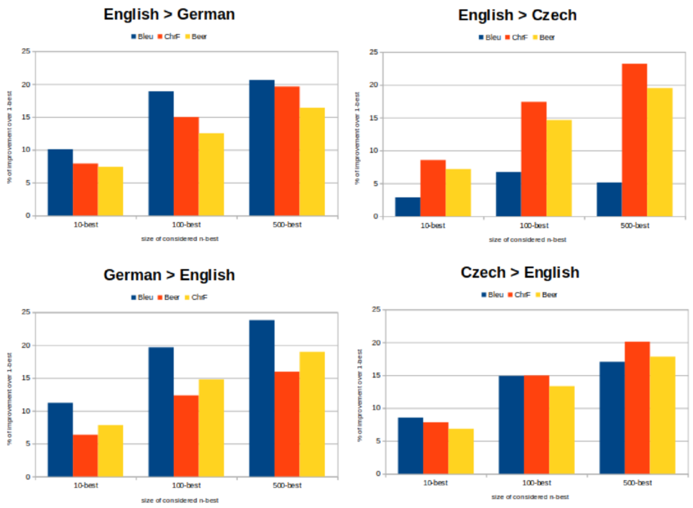

Figure 1: Proportion of improvement in NMT results according to MT evaluation metrics based on the oracle results over the original 1-best when the size of the beam is increased for decoding.

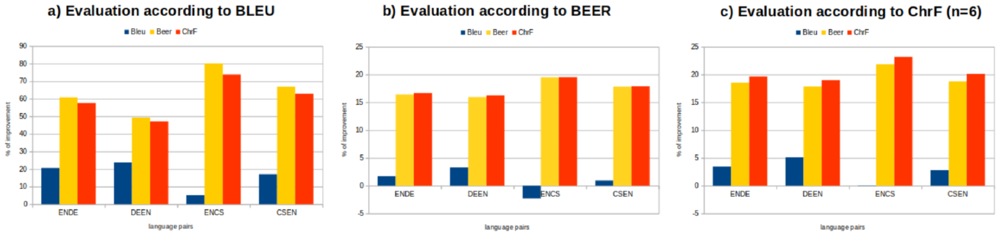

Figure 2: Focusing on the 500-best output for NMT, which brings the best gains in Figure 1, proportion of improvement of the oracle translation over the original 1-best when using different metrics for the oracle computation: ChrF3, BEER and BLEU. Re-ranking is done with one metric at a time, and the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

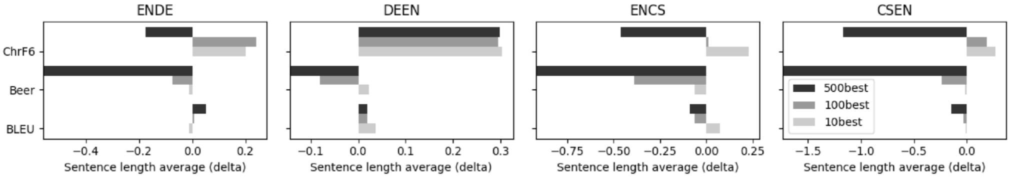

Figure 3: Delta in average sentence length for all NMT systems after 1-best oracle translation selection by each metric, compared to the average sentence length of the original 1-best.

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.