Neural Machine Translation
Leveraging Phrase-based Models in a Hybrid Search

Leonard Dahlmann and **Evgeny Matusov** and **Pavel Petrushkov** and **Shahram Khadivi**
eBay Inc.
Kasernenstr. 25
52064 Aachen, Germany
{fdahlmann, ematusov, ppetrushkov, skhadivi}@ebay.com

Abstract

In this paper, we introduce a hybrid search for attention-based neural machine translation (NMT). A target phrase learned with statistical MT models extends a hypothesis in the NMT beam search when the attention of the NMT model focuses on the source words translated by this phrase. Phrases added in this way are scored with the NMT model, but also with SMT features including phrase-level translation probabilities and a target language model. Experimental results on German→English news domain and English→Russian e-commerce domain translation tasks show that using phrase-based models in NMT search improves MT quality by up to 2.3% BLEU absolute as compared to a strong NMT baseline.

1 Introduction

Neural machine translation has become state-of-the-art in recent years, reaching higher translation quality than statistical phrase-based machine translation (PBMT) on many tasks. Human analysis (Bentivogli et al., 2016) showed that NMT makes significantly fewer reordering errors, and also is able to select correct word forms more often than PBMT in the case of morphologically rich target languages. Overall, the fluency of the MT output improves when NMT is used, and the number of lexical choice errors is also reduced. However, state-of-the-art NMT approaches based on an encoder-decoder architecture with an attention mechanism as introduced by (Bahdanau et al., 2014) exhibit weaknesses that sometimes lead to MT errors which a phrase-based MT system does not make. In particular, PBMT usually can better translate rare words (e.g. singletons), as well as memorize and use phrasal translations. NMT has problems translating rare words because of limitations on the vocabulary size, as well as the fact that word embeddings are used to represent both source and target words. A rare word's embedding can not be trained reliably.

Another handicap of NMT is a general difficulty of fixing errors made by a neural MT system. Since NMT does not explicitly use or save word-to-word or phrase-to-phrase mappings, and its search is a target word beam search with almost no constraints, it is difficult to fix errors by an NMT system. It is important to quickly fix certain errors in real-life applications of MT systems to avoid negative user feedback or other (e.g. legal) consequences. An error identified in the output of a PBMT system can be fixed by tracing which phrase pair was used that resulted in the error, and down-weighting or even removing the phrase pair. Also, in PBMT it is easy to add an "override" translation.

In this work, we combine the strengths of NMT and PBMT approaches by introducing a novel hybrid search algorithm. In this algorithm, the standard NMT beam search is extended with phrase translation hypotheses from a statistical phrase table. The decision on when to use what phrasal translations is taken based on the attention mechanism of the NMT model, which provides a soft coverage of the source sentence words. All partial phrasal translations are scored with the NMT decoder and can be continued with a word-based NMT translation candidate or another phrasal translation candidate.

The proposed search algorithm uses a log-linear model in which the NMT translation score is combined with standard phrase translation scores, including a target n-gram language model (LM) score. Thus, a LM trained on additional monolingual data can be used. The decisions on the word

order in the produced target translation are taken based only on the states of the NMT decoder.

This paper is structured as follows. We review related work in Section 1.1. The baseline NMT model we use is described in Section 2, where we also recap the log-linear model combination used in PBMT. Section 3 presents the details of the proposed hybrid search. Experimental results are presented in Section 4, followed by conclusions and outlook in Section 5.

1.1 Related Work

In the line of research closely related to our approach, neural models are used as additional features in vanilla phrase-based systems. Examples include the work of (Devlin et al., 2014), (Junczys-Dowmunt et al., 2016), etc. Such approaches have certain limitations: first, the search space of the model is still restricted by what can be produced using a phrase table extracted from parallel data based on word alignments. Second, the organization of the search, in which only a limited target word history (e.g. 4 last target words) is available for each partial hypothesis, makes it difficult to integrate recurrent neural network LMs and translation models which take all previously generated target words into account. That is why, for instance, the attention-based NMT models were usually applied only in rescoring (Peter et al., 2016).

In (Stahlberg et al., 2017), a two-step translation process is used, where in the first step a SMT translation lattice is generated, and in the second step the NMT decoder combines NMT scores with the Bayes-risk of the translations according to the lattice. In contrast, we explicitly use phrasal translations and language model scores in an integrated search.

In (Arthur et al., 2016), a statistical word lexicon is used to influence NMT hypotheses, also based on the attention mechanism. (Gülçehre et al., 2015) combine target n-gram LM scores with NMT scores to find the best translation. (He et al., 2016) also use a target LM, but add further SMT features such as word penalty and word lexica to the NMT beam search. To the best of our knowledge, no previous work extends the beam search with phrasal translation hypotheses of PBMT, like we propose in this paper.

In (Tang et al., 2016), the NMT decoder is modified to switch between using externally defined phrases and standard NMT word hypotheses. However, only one target phrase per source phrase is considered, and the reported improvements are significant only when manually selected phrase pairs (mostly for rare named entities) are used.

Somewhat related to our work is the concept of coverage-based NMT (Tu et al., 2016), where the model architecture is changed to explicitly account for source coverage. In our work, we use a standard NMT architecture, but track coverage with accumulated attention weights.

2 Background

2.1 Neural MT

Neural MT proposed by (Bahdanau et al., 2014) maximizes the conditional log-likelihood of the target sentence $E : e_1, \ldots, e_I$ given the source sentence $F : f_1, \ldots, f_J$:

$$H_D = -\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(E_n|F_n)$$

where (E_n, F_n) refers to the n-th training sentence pair in a dataset D, and N denotes the total number of sentence pairs in the training corpus. When using the encoder-decoder architecture by (Cho et al., 2014), the conditional probability can be written as:

$$p(e_1 \cdots e_I|f_1 \cdots f_J) = \prod_{i=1}^{I} p(e_i|e_{i-1} \cdots e_1, c)$$

with $p(e_i|e_{i-1} \cdots e_1, c) = g(s_i, e_{i-1}, c)$, where I is the length of the target sentence and J is the length of source sentence, c is a fixed-length vector to encode the source sentence, s_i is a hidden state of RNN at time step i, and $g(\cdot)$ is a nonlinear function to approximate the word probability. When the attention mechanism is used, the vector c in each sentence is replaced by a time-variant representation c_i that is a weighted summary over a sequence of annotations $(h_1, \ldots, h_J)$, and h_j contains information about the whole input sentence, but with a strong focus on the parts surrounding the j-th word (Bahdanau et al., 2014). Then, the context vector can be defined as:

$$c_i = \sum_{j}^{J} \alpha_{ij} h_j \quad \text{where} \quad \alpha_{ij} = \frac{exp(r_{ij})}{\sum_{j=1}^{J} exp(r_{ij})}.$$

Therefore, α_{ij} is normalized over all source positions j. Also, $r_{ij} = a(s_{i-1}, h_j)$ is the attention model used to calculate the log-likelihood of

aligning the i-th target word to the j-th source word.

2.2 Phrase-based MT

The log-linear model, as introduced in (Och and Ney, 2002), allows decomposing the translation probability $Pr(e_1^I|f_1^J)$ by using an arbitrary number of features $h_m(f_1^J, e_1^I)$. Each feature is multiplied by a corresponding scaling factor λ_m:

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(f_1^J, e_1^I)\right)}{\sum_{\tilde{e}_1^{\tilde{I}}} \exp\left(\sum_{m=1}^M \lambda_m h_m(f_1^J, \tilde{e}_1^{\tilde{I}})\right)}.$$

The standard PBMT approach uses a log-linear model in which bidirectional phrasal and lexical scores, language model scores, distortion scores, word penalties and phrase penalties are combined as features.

3 Hybrid Approach

In this section we describe our proposed hybrid NMT approach. The algorithm allows translations to be generated partially by phrases[1] and partially by words. Section 3.1 describes the models we use to score hypotheses. The search algorithm is presented in Section 3.2.

3.1 Log-linear Combination

We use a log-linear model combination to introduce SMT models into the NMT search. Since translations can be partially generated by phrases, we introduce the phrase segmentation s_1^K as a hidden variable into the models similarly to (Zens and Ney, 2008), where K is the number of phrases used in the translation. Note that, unlike standard PBMT, s_1^K does not need to cover the whole source sentence, as parts of the translation can be generated by words. Using the maximum approximation, the search criterion then is

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I} \left\{ \max_{s_1^K} \sum_{m=1}^M \lambda_m h_m(f_1^J, e_1^I, s_1^K) \right\}. \quad (1)$$

Let $\tilde{f}_k, \tilde{e}_k$ be the chosen phrase pairs in the segmentation s_1^K for $k = 1, \ldots, K$. In our experiments with the proposed hybrid search, we use the following features:

1. The NMT feature h_{NMT}.
2. The word penalty feature h_{WP} counts the number of target words. This feature can help control the length of translations.
3. The source word coverage feature h_{SWC} counts the number of source words translated by phrases:
$$h_{\text{SWC}}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^K |\tilde{f}_k|.$$
The purpose of this feature is to control the usage of phrases.
4. The phrase penalty feature h_{PP} counts the number of phrases used. Together with the word penalty and the source word coverage feature, the phrase penalty can control the length of chosen phrases.
5. The n-gram language model feature h_{LM}.
6. The bidirectional phrase features h_{Phr} and h_{iPhr}. Note that these features are only applied for those parts of the translation that are generated by phrases. The other parts get a phrase score of zero.

The scaling factors λ_m are tuned with minimum error rate training (MERT) (Och, 2003) on n-best lists of the development set.

3.2 Search

The algorithm is based on the beam search for NMT, which generates translations one word per time step in a left-to-right fashion. We modify this search to allow hypothesizing phrases in addition to normal word hypotheses. The phrases are suggested based on the neural attention, starting from the source position with the maximal current attention. We only suggest phrases if a source position is focused. We check that suggested phrases do not overlap with already translated source words by keeping track of the sum of attention in previous time steps for each source position. Thus, the problem of global reordering is left entirely to the NMT model and we follow the attention when hypothesizing phrases.

Hypotheses are scored by NMT and SMT models. The beam is divided into two parts of fixed size: the word beam and the phrase beam. The phrase beam is used to score target phrases which were hypothesized from an entry in a previous word beam. In order to score a target phrase consisting of k words with the NMT model, we use k time steps, allowing us to keep the efficiency of batched NMT scoring. Once a target phrase has been fully scored (and if the hypothesis has

[1] As in SMT, phrases can consist of only a single token.

not been pruned), the hypothesis is returned to the word beam. Both beams are generated and pruned independently in each time step.

The algorithm has some hyper-parameters that need to be set manually. First, we have the beam size N_p for phrase hypotheses and the beam size N_w for word hypotheses. Second, τ_{focus} is the minimum attention that needs to be on a source position to consider it for extending with a phrase translation candidate whose source phrase starts on that position. Third, τ_{cov} is the minimum sum of attention of a source position over previous time steps at which it is considered to be covered. We do not hypothesize phrases that overlap with covered positions.

In the following, we describe the search in detail. Let f_1^J be the source sentence. Before search, we run the standard phrase matching algorithm on the source sentence to retrieve the translation options $E(j, j')$ for source positions $1 \leq j < j' \leq J$ from a given SMT phrase table. With each hypothesis h, we associate the following items:

- $C(h, j)$ is the sum of the NMT attention to source position j involved in generating the target words of h. This can be considered as a soft coverage vector for h.
- $Q(h)$ is the partial log-linear score of h according to Equation 1.
- $E(h)$ is the n-gram target word history of h.
- If h is a phrase hypothesis with target phrase $\tilde{e}$, of which k words already have been scored by NMT, then $P(h) := (\tilde{e}, k)$ is the phrase state.

Also, each hypothesis is associated with its corresponding NMT hidden state. We initialize the beam to consist of an empty word hypothesis. Each step of the beam search proceeds as follows:

1. Let $B = [B_w, B_p]$ be the previous beam with word/phrase hypotheses, respectively. First, we generate the attention vector $\alpha_{h,j}$ and the distribution over target words $\hat{p}_h(e)$ for each hypothesis $h \in B$ and word e in the NMT target vocabulary V_T using the NMT model in batched scoring [2].
2. Initialize new beam $[B'_w, B'_p] = [\emptyset, \emptyset]$.
3. Generate new word hypotheses: find the maximal N_w pairs (h, e) with $h \in B_w$ and $e \in V_T$ according to the score $Q(h) + \lambda_{\text{NMT}} \cdot \log \hat{p}_h(e)$. For the top pairs $h' = (h, e)$, set

$$Q(h') = Q(h) + \lambda_{NMT} \cdot \log \hat{p}_h(e) + \lambda_{LM} \cdot \log p_{\text{LM}}(e|E(h)) + \lambda_{WP}$$

   and insert h' into B'_w. Update the soft coverage $C(h', j) = C(h, j) + \alpha_{h,j}$ for $1 \leq j \leq J$.
4. Generate new phrase hypotheses: for each previous word hypothesis $h \in B_w$, convert the soft attention $C(h, \cdot)$ into a binary coverage set C, such that $j \in C$ iff. $C(h, j) > \tau_{\text{cov}}$. Identify the current NMT focus as

$$\hat{j} = \underset{1 \leq j \leq J,\ \alpha_{h,j} > \tau_{\text{focus}}}{\arg\max} \alpha_{h,j}.$$

   If there is no such j with $\alpha_{h,j} > \tau_{\text{focus}}$, no phrase hypotheses are generated from h in this step. Otherwise, for each source phrase length l with $C \cap \{\hat{j}, \hat{j}+1, \ldots, \hat{j}+l-1\} = \emptyset$ and each target phrase $\tilde{e} \in E(\hat{j}, \hat{j}+l)$, create a new hypothesis $h' = (h, \tilde{e}_1)$ with the score

$$\begin{aligned} Q(h') &= Q(h) + \lambda_{\text{NMT}} \cdot \log \hat{p}_h(e_1) \\ &+ \lambda_{\text{LM}} \cdot \log p_{\text{LM}}(\tilde{e}|E(h)) + |\tilde{e}| \cdot \lambda_{\text{WP}} \\ &+ \lambda_{\text{PP}} + l \cdot \lambda_{\text{SWC}}. \end{aligned} \quad (2)$$

   Note that, in this step, the full target phrase is scored using the language model, while only the first target word is scored using NMT. Initialize the phrase state of h': $P(h') = (\tilde{e}, 1)$. As in step 3, update the soft coverage. If $|\tilde{e}| = 1$, insert h' into B'_w, otherwise insert into B'_p.
5. Advance previous phrase hypotheses: for each $h \in B_p$, with phrase state $P(h) = (\tilde{e}, k)$, score the $(k+1)$-th target word of $\tilde{e}$ using NMT, setting $h' = (h, \tilde{e}_{k+1})$ and

$$Q(h') = Q(h) + \lambda_{\text{NMT}} \cdot \log \hat{p}_h(\tilde{e}_{k+1}).$$

   As in step 3, update the soft coverage. Set the new phrase state as $P(h') = (\tilde{e}, k+1)$. If $k + 1 = |\tilde{e}|$, we are finished scoring the phrase and h' is inserted into B'_w. Otherwise, h' is inserted in B'_p.
6. Prune B'_w to N_w entries and B'_p to N_p entries according to $Q(\cdot)$.
7. Insert all hypotheses from the pruned B'_w and B'_p where the last word is the sentence end token into the set of finished hypotheses B_f.
8. $B := [B'_w, B'_p]$.

[2] If a target word e is not in V_T, set $\hat{p}_h(e) = \hat{p}_h(\text{UNK})$ where UNK is a special token denoting unknowns. Note that this almost never happens when using a word segmentation like BPE (Sennrich et al., 2016b).

Data set		WMT		E-commerce	
Language		German	English	English	Russian
Training	Sentences	5,597,491		2,919,406	
	Running words	129,083,315	134,469,297	46,715,319	45,305,268
	Full vocabulary	1,961,186	884,075	326,015	774,435
Dev	Sentences	2169 (WMT 15)		950	
	Running words	56,593	51,324	24,487	24,087
Test	Sentences	6002 (WMT 14 + 16)		1051 (item/product descriptions)	
	Running words	160,469	144,387	29,165	26,476

Table 1: Corpus statistics for the WMT German→English and e-commerce English→Russian MT tasks.

If phrase scores from a phrase table are to be included in the search, Equation 2 needs to be modified by adding $\lambda_{\text{Phr}} \log p(\tilde{f}|\tilde{e})$ and $\lambda_{\text{iPhr}} \log p(\tilde{e}|\tilde{f})$.

As in the pure NMT beam search, this procedure is repeated until either the last word of all hypotheses in a step is the sentence end token, or $2 \cdot J$ many beam steps have been performed. Finally, the best translation is chosen as the one in B_f with the highest score.

Note that the same target sequence can be generated with different phrasal segmentations. During search, if two hypotheses have the same full target history in a beam, we recombine them and discard the hypothesis with the lower score.

4 Experiments

We perform experiments comparing the translation quality of our hybrid approach to phrase-based and pure end-to-end NMT baselines. We present results on two tasks: an in-house English→Russian e-commerce task (translation of real product/item descriptions from an e-commerce site), and the WMT 2016 German→English task (news domain). The corpus statistics are shown in Table 1.

For the English→Russian task, the parallel training data consists of an in-domain part (ca. 5.5M running words) of product/item titles and descriptions and other e-commerce content. The rest is out-of-domain data (UN, subtitles, TAUS data collections, etc.) sampled to have significant n-gram overlap with the in-domain description data. Item descriptions are provided by private sellers and, like any user-generated content, may contain ungrammatical sentences, spelling errors, and other noise. Product descriptions usually originate from product catalogs and are more "clean", but on the other hand, are difficult to translate because of rare domain-specific terminology. Both types of text contain itemizations, measurement units, and other structures which are usually not found in normal sentences. We tune the system on a development set that is a mix of product and item descriptions, and evaluate on separate product/item description test sets. For development and test sets, two reference translations are used.

The German→English system is trained on parallel corpora provided for the constrained WMT 2017 evaluation (Europarl, Common Crawl, and others). We use the WMT 2015 evaluation data as development set, and the evaluation is performed on two sets from the WMT evaluations in 2014 and 2016. Only a single human reference translation is provided.

For the phrase-based baselines, we use an in-house phrase-decoder (Matusov and Köprü, 2010) which is similar to the Moses decoder (Koehn et al., 2007). We use standard SMT features, including word-level and phrase-level translation probabilities, the distortion model, 5-gram LMs, and a 7-gram joint translation and reordering model reimplemented based on the work of (Guta et al., 2015). The language model for the e-commerce task is trained on additional monolingual Russian item description data containing 28.2M words. For the WMT task, we use the English News Crawl data containing 3.8B words for additional language model data. The tuning is performed using MERT (Och, 2003) to increase the BLEU score on the development set. To stabilize the optimization on the English→Russian task, we detach Russian morphological suffixes from the word stems both in hypotheses and references using a context-independent "poor man's" morphological analysis. We prefix each suffix with a special symbol and treat them as separate tokens.

We have implemented our NMT model in

		Item descriptions		Product descriptions	
System description	Beam size	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Phrase-based	-	21.3	61.6	22.7	56.6
+ 1000-best rescoring with NMT	-	23.1	60.1	25.8	54.7
NMT	12	26.4	56.4	28.4	52.0
NMT	128	26.3	56.6	28.5	51.9
Full hybrid approach	128	26.7	56.1	29.9	51.2
+ extra LM data	128	27.4	55.4	30.8	50.5
NMT + WP + LM (with extra data)	128	26.2	57.3	29.0	51.8

Table 2: Overview of translation results on the e-commerce English→Russian task.

Python using the TensorFlow[3] deep learning library. We use the embedding size of 620, RNN size of 1000 and GRU cells. The model is trained with maximum likelihood loss for 15 epochs using Adam optimizer (Kingma and Ba, 2014) on complete data in batches of 100 sentences. The learning rate is initialized to 0.0002, decaying by 0.9 each epoch. For regularization we use L2 loss with weight 10^{-7} and dropout following Gal and Ghahramani (2016). We set the dropout probability for input and recurrent connections of the RNN to 0.2 and word embedding dropout probability to 0.1. On the English→Russian task, the model is then fine-tuned on in-domain data for 10 epochs. The vocabulary is limited using byte pair encoding (BPE) (Sennrich et al., 2016b) with 40K splits separately for each language. To speed up training we use approximate loss as described in (Jean et al., 2015). For pure NMT experiments, we employ length normalization (Wu et al., 2016), as otherwise short translations would be favored.

For the hybrid approach, we use the same trained end-to-end model as in the NMT baseline. We use all the phrase-based model features plus the NMT score and run MERT as described in Section 3.1. Language models are trained on the level of BPE tokens. We consider at most 100 translation options for each source phrase. If not specified otherwise, we use a beam size of 96 for phrase hypotheses and a beam size of 32 for word hypotheses, resulting in a combined beam size of 128. Furthermore, we set the focus threshold $\tau_{\text{focus}} = 0.3$ and the coverage threshold $\tau_{\text{cov}} = 0.7$ by default. We also perform experiments where these hyper-parameters are varied.

[3] http://tensorflow.org

4.1 E-commerce English→Russian

The results on the e-commerce English→Russian task are summarized in Table 2.

NMT vs. phrase-based SMT

The pure NMT system exhibits large improvements over the phrase-based baseline[4]. These improvements are also significantly larger than when we use the NMT model to rescore PBMT 1000-best lists. NMT results are not improved when the beam size is increased from 12 to 128.

Hybrid search vs. pure NMT search

For the hybrid approach, we train a phrase-table on the in-domain data and split the source and target phrases with BPE afterwards for compatibility with the NMT vocabulary. With the hybrid approach, when using a LM trained only on the target side of bilingual data, we get an improvement of 0.3% BLEU on item descriptions and 1.4% BLEU on product descriptions over the pure NMT system. When we use the LM trained on extra monolingual data, we get total improvements of 1.0% BLEU and 2.3% BLEU with the hybrid approach. In contrast, when we add this language model and a word penalty on top of the pure NMT system and tune scaling factors with MERT, we get small improvements (last row of Table 2) only on product descriptions. This shows that the hybrid approach can exploit the LM better than a purely word-based NMT approach. We have also performed experiments utilizing the additional monolingual data for synthetic training data for NMT as in (Sennrich et al., 2016a), but did not get improvements.

To analyze the improvements of the hybrid system, we perform experiments in which we either

[4] The significance of these improvements was also confirmed by an in-house human evaluation with 3 judges.

	Item descriptions		Product descriptions	
System description	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Full hybrid approach	27.4	55.4	30.8	50.5
Without LM	26.5	55.9	29.2	51.0
Without source word coverage feature	26.7	56.1	29.4	51.2
Without phrase scores	27.2	55.9	30.6	50.6
Maximal source phrase length 1	26.7	56.4	29.1	51.6
Minimal source phrase length 2	27.0	55.9	30.0	51.1

Table 3: Translation results of the hybrid approach on the e-commerce English→Russian task with different SMT model combinations. The first row shows results with all models enabled. In the following rows, we either remove or limit exactly one model compared to the full system.

disable or limit some of the SMT models. The results are shown in Table 3. Without the language model, the hybrid approach has almost no improvements over the NMT baseline. This indicates that the language model is crucial in selecting appropriate phrase candidates. Similarly, when we disable the source word coverage feature, the translation quality is degraded, suggesting that this feature helps choose between phrase hypotheses and word hypotheses during the search. Next, we do not use phrase-level scores. Here, we observe only a small degradation of translation quality. Finally, we limit the source length of phrases used in the search, allowing only one-word source phrases in one experiment and only source phrases with two or more words in another experiment. In both cases, the translation quality decreases. Thus, both one-word phrases and longer phrases are necessary to obtain the best results.

Tuning the beam size

Next, we study the effect of different beam sizes on translation quality. The results are shown in Table 4. Note that we retune the system for each choice. With a total beam size of 128, we get the best results by using a phrase beam size of 96 and a word beam size of 32. When we use a phrase beam size of 116 or 64 instead, the translation quality worsens. In another experiment, we decrease the total beam size to 64. The translation quality degrades only slightly, which means that we can still expect MT quality improvements with hybrid search even if we optimize the system for speed. To further test this, we reduce the beam sizes to $N_w = 12$ and $N_p = 4$ after tuning with $N_w = 32$ and $N_p = 96$. We get BLEU scores of 27.1% on item descriptions and 30.1% on product descriptions, losing 0.3% and 0.7% BLEU respectively compared to the full beam size.

Beam size		Item descr.		Product descr.	
N_p	N_w	BLEU [%]	TER [%]	BLEU [%]	TER [%]
116	12	26.7	55.9	29.8	51.1
96	32	27.4	55.4	30.8	50.5
64	64	26.8	55.6	30.1	50.7
32	32	27.1	55.8	30.7	50.5

Table 4: Effect of the beam size (word beam size N_w + phrase beam size N_p) for the hybrid approach on the e-commerce English→Russian task.

Tuning the attention focus/coverage thresholds

Table 5 shows results with different values for the coverage threshold τ_{cov}. Again, we retune the system for each choice. Setting the coverage threshold to 1.0 or even disabling the coverage check (by setting $\tau_{\text{cov}} = \infty$) has little effect on the translation scores on this task. This can be explained by the fact that translation from English to Russian is mostly monotonic. We also tried varying the focus threshold τ_{focus} between 0.0 and 0.3 but did not notice any significant effect on this task.

		Item descr.		Product descr.	
τ_{focus}	τ_{cov}	BLEU [%]	TER [%]	BLEU [%]	TER [%]
0.3	0.7	27.4	55.4	30.8	50.5
0.3	1.0	27.2	55.4	30.3	50.3
0.3	∞	27.5	55.4	30.4	50.9

Table 5: Effect of the threshold parameters on the hybrid approach on the e-commerce English→Russian task.

Analysis

To understand the behavior of the hybrid search, we count the number of source words that are

		newstest2014		newstest2016	
System description	Beam size	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Phrase-based	-	22.9	59.4	26.9	54.1
+ News Crawl LM data	-	25.4	59.0	29.2	53.8
NMT	12	26.9	53.0	32.3	47.6
NMT	64	27.0	53.0	32.2	47.6
Hybrid approach	64	27.8	53.2	32.4	48.2
+ tuning τ_{focus}, τ_{cov}	64	28.0	53.0	33.3	47.4
+ News Crawl LM data	64	29.7	52.2	35.3	46.7

Table 6: Overview of translation results on the WMT German→English task.

translated by phrases in the product descriptions test set. Of the 9320 source words, 7109 (76.3%) are covered by phrase hypotheses. 78.3% of the source phrases are unigrams, 19.5% are bigrams and 2.2% are trigrams or longer. Among the many one-word phrases used, almost all (99.2%) are also within the top 3 predictions of word-based NMT, and 90.3% are equal to the top NMT prediction.

Further human analysis by a native Russian speaker of the pure NMT vs. hybrid search translations shows that hybrid search is often able to correct the following known NMT handicaps:

- incorrect translation of rare words (among other reasons, due to incorrect sub-word unit translation in which rare words are aggressively segmented).
- repetition of same or similar words as a result of multiple attention to the same source word, as well as untranslated words that received no attention.
- incorrect or partially correct word-by-word translation when a phrasal (non-literal) translation should be used instead.

In all of these cases, the usage of phrasal translations is able to better enforce the coverage, and this, in turn, leads to improved lexical choice. The fact that not many long phrase pairs are selected indicates, in our opinion, that the search and modeling problem in NMT is far from being solved: with the right, diverse model scores, the proposed hybrid search is able to select and extend better hypotheses with words, most of which already had a high NMT probability. Yet they are not always selected in the pure NMT beam search, among other reasons, due to competition from words erroneously placed near them in the embedding space.

4.2 WMT 2016 German→English

The results on the WMT German→English task are shown in Table 6. The initial phrase-based baseline uses the 5-gram language model estimated on the target side of bilingual data. By adding the News Crawl LM data, we gain 2.5% and 2.3% BLEU on the test sets, but PBMT still is behind NMT.

For the hybrid approach, we use a beam size of 64 and a maximal number of beam steps of $1.5 \cdot J$ (instead of $2 \cdot J$) to speed up experiments. We use separate word penalty features, one for word-based hypotheses and one for phrase-based hypotheses to allow for more control of translation lengths. With the hybrid approach, using the 5-gram language model estimated on the target side of bilingual data, and phrase scores, we get small improvements in BLEU over the NMT baseline. However, the TER increases. We experiment with different thresholds, setting $\tau_{\text{focus}} = 0.1$ and $\tau_{\text{cov}} = 1.0$. With this hybrid system, we get improvements of 1.0% and 1.1% BLEU over pure NMT. Finally, we add the News Crawl LM data on top. This significantly improves the results by 1.7% and 2.0% BLEU. In total, we gain 2.7% and 3.1% BLEU over pure NMT. These results reinforce the fact that, similar to PBMT, language model quality is important for the proposed hybrid search. In contrast, we have also tried applying only the LM (including News Crawl data) with a word penalty on top of NMT, but did not get consistent improvements.

Figure 1 shows an example for the phrase pairs chosen by the hybrid system on top of the NMT attention. The hybrid approach correctly translates the German idiom “nach und nach” as “gradually”, while the pure NMT system incorrectly translates it word-by-word as “after and after”.

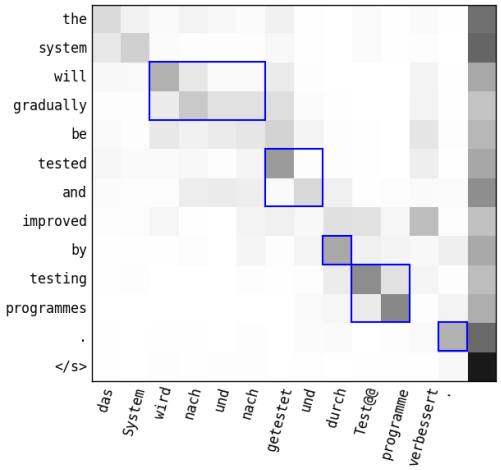

Figure 1: Example alignment from the hybrid search, with the source sentence on the bottom and the translation on the left. The blue rectangles signify phrase pairs on top of the NMT attention. The pure NMT translation is "the system is tested after and after testing and improved by testing programs."

5 Conclusion

In this work, we proposed a novel hybrid search that extends NMT with phrase-based models. The NMT beam search was modified to insert phrasal translations based on the current and accumulated attention weights of the NMT decoder RNN. The NMT model score was used in a log-linear model with standard phrase-based scores as well as an n-gram language model. We described the algorithm in detail, in which we keep separate beams for NMT word hypotheses and hypotheses with an incomplete phrasal translation, as well as introduce parameters which control the source sentence coverage. Numerous experiments on two large vocabulary translation tasks showed that the hybrid search improves BLEU scores significantly as compared to a strong NMT baseline that already outperforms phrase-based SMT by a large margin.

In the future, we plan to focus on integration of phrasal components into NMT training, including better coverage constraints, as well as methods for context-dependent translation override within our hybrid search algorithm.

Acknowledgments

We would like to thank Tamer Alkhouli and Jan-Thorsten Peter for helpful discussions. We thank the anonymous reviewers for their suggestions.

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1557–1567.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 257–267.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1724–1734.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*. pages 1370–1380.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*. pages 1019–1027.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR* abs/1503.03535.

Andreas Guta, Tamer Alkhouli, Jan-Thorsten Peter, Joern Wuebker, and Hermann Ney. 2015. A comparison between count and neural network models based on joint translation and reordering sequences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1401–1411.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. pages 151–157.

Sébastien Jean, KyungHyun Cho, Roland Memisevic, and Yoshua Bengio. 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 of the Asian Federation of Natural Language Processing. pages 1–10.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL*. pages 319–325.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Evgeny Matusov and Selçuk Köprü. 2010. AppTek's APT Machine Translation System for IWSLT 2010. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *International Workshop on Spoken Language Translation, IWSLT*. pages 29–36.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. pages 160–167.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. pages 295–302.

Jan-Thorsten Peter, Andreas Guta, Nick Rossenbach, Miguel Graa, and Hermann Ney. 2016. The rwth aachen machine translation system for iwslt 2016. In *International Workshop on Spoken Language Translation, IWSLT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 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*. Valencia, Spain.

Yaohua Tang, Fandong Meng, Zhengdong Lu, Hang Li, and Philip L. H. Yu. 2016. Neural machine translation with external phrase memory. *CoRR* abs/1606.01792.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144.

Richard Zens and Hermann Ney. 2008. Improvements in dynamic programming beam search for phrase-based statistical machine translation. In *International Workshop on Spoken Language Translation, IWSLT*. pages 198–205.