Multi-Pass Decoding With Complex Feature Guidance for Statistical Machine Translation

Benjamin Marie
LIMSI-CNRS, Orsay, France
Lingua et Machina, Le Chesnay, France
benjamin.marie@limsi.fr

Aurélien Max
LIMSI-CNRS, Orsay, France
Univ. Paris Sud, Orsay, France
aurelien.max@limsi.fr

Abstract

In Statistical Machine Translation, some complex features are still difficult to integrate during decoding and usually used through the reranking of the k-best hypotheses produced by the decoder. We propose a translation table partitioning method that exploits the result of this reranking to iteratively guide the decoder in order to produce a new k-best list more relevant to some complex features. We report experiments on two translation domains and two translations directions which yield improvements of up to 1.4 BLEU over the reranking baseline using the same set of complex features. On a practical viewpoint, our approach allows SMT system developers to easily integrate complex features into decoding rather than being limited to their use in one-time k-best list reranking.

1 Introduction

State-of-the-art Phrase-Based Statistical Machine Translation (PBSMT) systems can use a large number of feature functions decomposable into local scores to efficiently evaluate the partial hypotheses built during decoding. However, some feature functions are difficult to integrate into the decoder mainly because they are not easily decomposable, very costly to compute and/or only available after complete hypotheses have been posited. Usually such *complex features* are used through the rescoring and reranking of the k-best translation hypotheses produced by the decoder (Och et al., 2004). Although this reranking pass is performed over the best part of the decoder search space, it is limited by the actual *diversity* expressed in the k-best list. Additionally, reranking being performed on a list generated by a simpler set of features, it may not have access to hypotheses that can best exploit the potential of the complex features used. We describe a translation table partitioning approach that exploits the result of such a reranking to iteratively guide the decoder to produce new hypotheses that are more relevant to the complex features used. To this end, we focus in this work on the simple exploitation of the disagreement between hypotheses ranked best according to the decoder and to our feature-rich decoder. In particular, we seek to provide the next-pass decoder with separate translation tables that either contain bi-phrases that are unique to the decoder's one-best or to the reranker's one-best, in the hope that it will tend, in a soft manner, to *exploit* the preferences expressed by the complex features, and to otherwise *explore* alternative translation choices. Such a comparison is then iteratively repeated, until convergence on a development set between the new pass of the decoder and a reranker trained on the full set of hypotheses generated thus far. On the test data, this procedure thus produces after each iteration a new decoder n-best, as well as an iteration-specific new reranker best hypothesis. We report consistent improvements of translation quality over a strong reranking baseline using the same features on 2 different domains and 2 translation directions.

The remainder of this article is organized as follows: we first briefly review related work (Section 2), then introduce our approach (Section 3), describe our experiments (Section 4), and finally discuss our results and present our future work (Section 5).

2 Related Work

Chen et al. (2008a; 2008b) expand the k-best list of the decoder using three methods. One of them involves re-decodings using models trained on the decoder k-best list to integrate posterior knowledge during the next re-decoding. The new k-best

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 554–559,
Beijing, China, July 26-31, 2015.

list produced by the decoder is concatenated to the original one and then reranked with complex features, which yields improvements over a reranking performed on the original k-best list. The reranking pass is done out of the loop and the redecodings do not exploit the reranking result that used the complex features.

Recently, we proposed a rewriting system that explores in a greedy fashion the neighborhood of the one-best hypothesis found by the reranking pass using complex features, assuming that a better hypothesis can be very close to this seed hypothesis (Marie and Max, 2014). Nevertheless, this rewriting only explores a small search space, limited by the greedy search algorithm that concentrates on individual, local rewritings.

Other works proposed methods to produce more diverse lists of hypotheses by iteratively encouraging the decoder to produce translations that are different from the previous one (Gimpel et al., 2013) or by making small changes to the scoring function to extract k-best lists from other parts of the search space (Devlin and Matsoukas, 2012). Some useful diversity can be obtained as these hypotheses can be combined using SMT system combination or help to better train reranking systems. But in spite of the introduction of more diversity, these methods do not guarantee that eventually lists containing hypotheses that are more relevant to complex features will be obtained.

3 Translation Table Partitioning

3.1 Exploiting the Reranking Pass Result

Because all bi-phrases initially belong to the same translation table, they share their feature weights after tuning. Our main idea is to partition the set of bi-phrases by putting aside, in new translation tables, possibly misused bi-phrases according to the reranking with complex features of the decoder k-best list (Rerank). This partitioning gives to subsequent tunings the opportunity to assign more adapted weights to the features of these specific groups of bi-phrases. Intuitively, if the Rerank one-best hypothesis is different from that of the initial decoder, the bi-phrases that account for the differences should have received different weights to encourage the decoder to either choose them or instead avoid them.

To achieve the partitioning of the translation table we compare the Rerank one-best hypothesis to the decoder one-best and compute their differences. On the one hand, there are n-grams from the decoder one-best hypothesis that are not found any more in the Rerank one-best; on the other hand, there are n-grams that only exist in the Rerank one-best hypothesis. Since the decoder produces word alignments between the source sentence to translate and its hypotheses, we can extract all the bi-phrases from the translation table that are compatible with these n-grams and their alignments. Each set of bi-phrases extracted from n-grams[1] either appearing (IN) or disappearing (OUT) in the Rerank one-best hypothesis compared to the decoder's, is moved to a specific translation table. Then a new tuning is performed for each relevant partitioning configuration.

The described translation table partioning procedure can be performed iteratively as each new decoding can be followed by Rerank on the new k-best list generated. The differences between Rerank and the decoder one-bests are extracted anew and put in new translation tables at each iteration.[2] Iterations are performed until no more improvements of the BLEU score are obtained by Rerank on a development set. The decoder is retuned and Rerank is re-trained after each iteration[3] to obtain more specific and updated weights for each old or new translation table. Finally, at test time, the learned weights corresponding to the current iteration are applied.

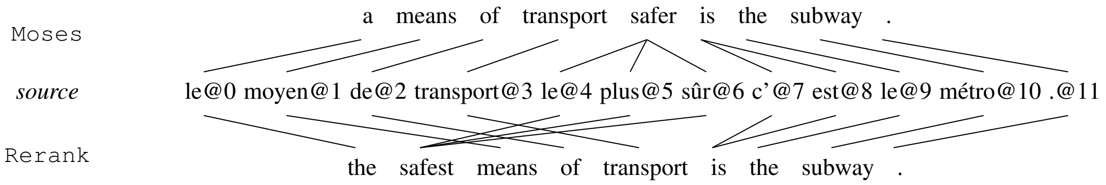
3.2 Located Tokens

As a token can appear more than once in an input text and in a sentence, and because complex features are computed locally, the source tokens are *located*: an identifier is concatenated to each token to make them unique in the source text to translate. Tokens of source phrases in the translation table are also located, meaning that each bi-phrases is duplicated to cover all located tokens. This procedure allows our approach to differentiate changes between Moses and Rerank one-best hypotheses at the token level by taking context into ac-

[1] In decoders phrases typically have a fixed maximum length, which corresponds to our maximum value for n.

[2] So, if both types of translation tables are extracted at each iteration, 3 iterations would produce 6 translation tables in addition to the remainder of the initial one. Note that a bi-phrase can in fact be present in more than one translation table after several iterations.

[3] Rerank re-training uses only the k-best list of the current iteration. k-bests from different iteration cannot be concatenated as they use a different number of features corresponding to a different number of translation tables.


Moses
a means of transport safer is the subway .
source
le@0 moyen@1 de@2 transport@3 le@4 plus@5 sûr@6 c'@7 est@8 le@9 métro@10 .@11
Rerank
the safest means of transport is the subway .

source	OUT	IN
le@0	a	the
le@0 moyen@1	a means	
moyen@1 de@2 transport@3	a means of transport	
le@4 plus@5 sûr@6	safer	safest

Figure 1: Example of IN and OUT translation tables extraction from the n-grams that differ between the Rerank and Moses one-best hypotheses.

count. An example of IN and OUT translation tables extraction with located tokens is presented in Figure 1.

4 Experiments

4.1 Data

We ran experiments on two translation tasks for different domains: the WMT'14 Medical translation task (medical) and the WMT'11 news translation task (news) for the language pair Fr-En on both directions. For both tasks we trained two strong baseline systems using data provided by WMT[4]. Statistics about the training, development and testing data are presented in Table 1.

Tasks	Corpus	Sentences	Tokens (Fr-En)
news	train	12M	383M - 318M
	dev	2,525	73k - 65k
	test	3,003	85k - 74k
medical	train	4.9M	91M - 78M
	dev	500	12k - 10k
	test	1,000	26k - 21k
	in-domain LM		146M - 78M
for both tasks	LM		2.5B - 6B

Table 1: Data used in our experiments.

4.2 MT system

For our experiments we used the Moses phrase-based SMT toolkit (Koehn et al., 2007) with default settings and features, including the five features from the translation table, and kb-mira tuning (Cherry and Foster, 2012). Rerank is trained using kb-mira on the 1,000-best list generated by Moses on the development set with the distinct-nbest parameter to have no duplicates. Testing is also performed on distinct 1,000-best lists. Rerank uses all the decoder features along with the following complex features:

- **MosesNorm**: all decoder features and the Moses score normalized by the hypothesis length
- **NNM**: bilingual and monolingual neural network models with a structured output layer (SOUL) (Le et al., 2012)
- **POSLM**: 6-gram POS language model
- **WPP**: count-based word posterior probability (Ueffing and Ney, 2007)
- **TagRatio**: ratio of translation hypothesis by number of source tokens tagged as: verb, noun or adjective
- **Syntax**: depth, number of nodes and number of unary rules of the syntactic parse normalized by the hypothesis length (Carter and Monz, 2011)
- **IBM1**: IBM1 features (Och et al., 2004; Hildebrand and Vogel, 2008)

Part-of-speech tagging and syntactic parsing were respectively performed with the Stanford Part-of-speech Tagger (Toutanova and Manning, 2000) and the Shift-Reduce parser of Zhu *et al.* (2013). We report the individual performance of each feature set in Table 2 and the Rerank performance when using all feature sets. As expected, the **NNM** feature set brings most of the improvements and attain by itself nearly the BLEU score of Rerank when using all feature sets for the news task with a gain of 1.4 and 1.1 BLEU respectively for En→Fr and Fr→En over the Moses

[4] http://www.statmt.org/wmt14

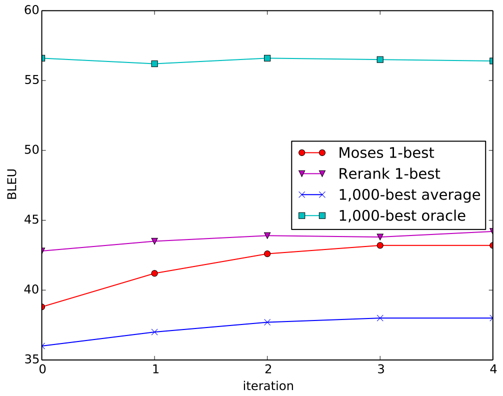


Figure 2: BLEU score evolution over iterations for the IN configuration on the test set of the medical En→Fr translation task.

baseline. Among the other feature sets, **POSLM** performs well, especially for the medical task with an improvement of 0.3 and 0.5 BLEU for En→Fr and Fr→En, respectively.

Some types of our complex features have already been used during decoding, although sometimes for a very important cost (Schwartz et al., 2011). Our feature sets are to be considered only as experimental parameters, as any other feature types usually used during reranking could also be used.

Features	medical		news	
	En→Fr	Fr→En	En→Fr	Fr→En
Moses	38.8	37.1	31.1	28.6
+ **MosesNorm**	38.9	37.2	31.1	28.7
+ **NNM**	41.9	38.9	32.5	29.8
+ **POSLM**	39.2	37.7	31.1	28.9
+ **WPP**	39.1	37.1	31.2	28.6
+ **TagRatio**	38.9	37.3	31.1	28.8
+ **Syntax**	38.8	37.2	31.2	28.9
+ **IBM1**	39.1	37.2	30.9	28.8
Rerank	42.8	40.1	32.5	29.9

Table 2: Reranking results for each set of features added individually; Rerank uses the full set.

4.3 Results

Table 3 presents our results for different translation table partitioning configurations. For each configuration, results are presented for the last iteration of the multi-pass decoding performed by Moses and the reranking of its k-best list by the Rerank system using complex features. First, we observe for the baseline systems that Rerank outperforms Moses for all translation tasks and directions, especially on medical with improvements of 3.0 and 4.0 BLEU respectively for Fr→En and En→Fr. These improvements illustrate the strong potential of our set of complex features to provide more accurate scores for translation hypotheses than the set of features used during the initial decoding.

All studied configurations yield improvements with multi-pass Moses over the Moses baseline, showing the advantage of extracting from the main translation table misused bi-phrases according to a reranking pass done with complex features. As illustrated by Figure 2, the multi-pass decoding quickly reduces the gap in BLEU score between our multi-pass Moses and Rerank one-best hypotheses. Although the 1,000-best oracle remains at the same level over the iterations, the 1,000-best average score[5] increases by 2 BLEU at the last iteration over the first 1,000-best hypotheses produced by Moses, pointing out a strong improvement of the average quality of the 1,000-best hypotheses. However, except for the IN configuration on medical En→Fr, multi-pass Moses does not bring improvements by itself over the Rerank baseline. Nevertheless, multi-pass Moses coupled with Rerank does improve over Rerank baseline for all configurations on all translation tasks. These consistent improvements over the Rerank baseline demonstrate the ability of our procedure to help the Moses decoder to produce k-best lists of better quality which are more suitable to our complex features.

The IN configuration, which puts in a translation table all bi-phrases in the one-best hypothesis of Rerank that do not belong to the Moses one-best hypothesis, performs the best for all translation tasks: multi-pass Rerank yields a 1.4 BLEU improvement over the Rerank baseline on medical En→Fr, and 0.7 BLEU on news En→Fr. Improvements are lower, but nonetheless consistent, for the Fr→En direction, with +0.9 and +0.5 BLEU respectively on the medical and news tasks. The OUT configuration yields smaller improvements in comparison, meaning that putting aside (a few) first-ranked bi-phrases downgraded by Rerank is less useful in order to produce better k-best lists with Moses. Using in the same system both IN and

[5] To obtain this average we compute the arithmetic mean of the 1,000-best hypotheses sentence-BLEU scores and select the hypothesis with the closest score to the mean. Once we have selected an hypothesis for each sentence, the BLEU score is computed.

Configuration		medical En→Fr			medical Fr→En			news En→Fr			news Fr→En		
		dev	test	# iter.	dev	test	# iter.	dev	test	# iter.	dev	test	# iter.
baseline	Moses	40.9	38.8	-	41.3	37.1	-	27.1	31.1	-	28.0	28.6	-
	Rerank	43.9	42.8		44.2	40.1		28.5	32.5		29.1	29.9	
OUT	Moses	43.3	41.8	4	43.0	38.7	3	27.9	31.8	1	28.5	29.2	1
	Rerank	45.3	43.8		44.5	40.5		28.5	32.9		29.2	30.3	
IN	Moses	45.1	43.2	4	43.6	39.9	3	28.4	32.4	2	28.6	29.3	2
	Rerank	45.7	**44.2**		45.0	**41.0**		28.8	**33.2**		29.3	**30.4**	
IN and OUT	Moses	44.8	42.4	4	42.8	38.7	3	28.3	32.1	2	28.8	29.2	2
	Rerank	45.3	43.5		44.5	40.6		28.7	32.9		29.3	**30.4**	

Table 3: Results for different translation table partitioning configurations. OUT: configuration with a translation table containing bi-phrases of the Moses 1-best not in the Rerank 1-best. IN: configuration with a translation table containing bi-phrases of the Rerank 1-best not in the Moses 1-best. For all configuration the main translation table is still used but does not contain the extracted bi-phrases.

OUT iteration-specific translation tables ("IN and OUT") yields a performance situated between using IN and OUT separately, but which still consistently improves over the baseline Rerank.

5 Discussion and future work

We have presented a method for guiding a phrase-based decoder with translation tables partitioned on the basis of k-best list reranking making use of complex features. Our results showed consistent improvements in BLEU score over a strong Rerank baseline using the same features. We experimented with a simple criterion for iteratively partitioning the original phrase table of the system, and found that focusing on providing the next iteration decoder with the bi-phrases that were prefered at first rank by Rerank (IN) performed best.[6]

We now intend to study how to better take advantage of the expected characteristics of our IN and OUT tables, possibly by adding more features to our iteration-specific tables, or by exploiting information on bi-phrases computed on the full reranked lists. For our future work, we also plan to study approaches that can enhance the *diversity* in the k-best lists (Chatterjee and Cancedda, 2010; Gimpel et al., 2013) between each iteration of the multi-pass decoding to train a better Rerank after each decoding pass. Another area for improvement lies in the addition of yet more complex features, for instance to allow a better discourse coherence modelling over iterations (Ture et al., 2012; Hardmeier et al., 2012). Going further, we could study the effect of using other hypotheses instead of the Rerank one-best to perform the comparison with the Moses one-best hypothesis. For instance, we can reasonably expect that making this comparison with the output of a rewriting system, such as the one proposed in our previous work (Marie and Max, 2014), could extract more misused and useful bi-phrases on which to base our translation table partitioning since this rewriting system's output is usually better than the Rerank one-best and not in the k-best list of the decoder.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments, Quoc Khanh Do for his assistance on using SOUL, and Li Gong and Nicolas Pécheux for providing the authors with data used in the experiments. The work of the first author is supported by a CIFRE grant from French ANRT.

References

Simon Carter and Christof Monz. 2011. Syntactic Discriminative Language Model Rerankers for Statistical Machine Translation. *Machine Translation*, 25:317–339.

Samidh Chatterjee and Nicola Cancedda. 2010. Minimum error rate training by sampling the translation lattice. In *Proceedings of EMNLP*, Cambridge, USA.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008a. Exploiting N-best Hypotheses for SMT Self-

[6] Interestingly, a control experiment showed that using iteration-specific tables yields slightly better performance than fusioning all bi-phrases of a given type in a non iteration-specific table, possibly allowing later tunings to prefer the contents of the most recent, and possibly more reliable tables.

Enhancement. In *Proceedings of ACL, short papers*, Columbus, USA.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008b. Regenerating Hypotheses for Statistical Machine Translation. In *Proceedings of COLING*, Manchester, UK.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of NAACL*, Montréal, Canada.

Jacob Devlin and Spyros Matsoukas. 2012. Trait-based Hypothesis Selection for Machine Translation. In *Proceedings of NAACL*, Montréal, Canada.

Kevin Gimpel, Dhruv Batra, Chris Dyer, Gregory Shakhnarovich, and Virginia Tech. 2013. A Systematic Exploration of Diversity in Machine Translation. In *Proceedings of EMNLP*, Seatlle, USA.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of EMNLP-CoNLL*, Jeju Island, Korea.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of AMTA*, Honolulu, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous Space Translation Models with Neural Networks. In *Proceedings of NAACL*, Montréal, Canada.

Benjamin Marie and Aurélien Max. 2014. Confidence-based Rewriting of Machine Translation Output. In *Proceedings of EMNLP*, Doha, Qatar.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of NAACL*, Boston, USA.

Lane Schwartz, Chris Callison-Burch, William Schuler, and Stephen Wu. 2011. Incremental syntactic language models for phrase-based translation. In *Proceedings of ACL*, Portland, USA.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-speech Tagger. In *Proceedings of EMNLP*, Hong Kong.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of NAACL*, Montréal, Canada.

Nicola Ueffing and Hermann Ney. 2007. Word-Level Confidence Estimation for Machine Translation. *Computational Linguistics*, 33(1):9–40.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and Accurate Shift-Reduce Constituent Parsing. In *Proceedings of ACL*, Sofia, Bulgaria.