Better N-best Translations through Generative n-gram Language Models

Boxing Chen, Marcello Federico, Mauro Cettolo

FBK-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050 Povo di Trento, Italy
{boxing,federico,cettolo}@itc.it
http://hermes.itc.it

Abstract

Most statistical machine translation systems take advantage of a re-scoring step that is applied on a set of highly probable translation hypotheses computed by a decoding algorithm. Assuming the standard log-linear model framework, re-scoring usually improves over the most likely translation by the decoder because better and more sophisticated feature functions can be deployed. Clearly, this happens under the assumption that better translations exist indeed among the N-best ones. In this paper we present a technique to expand existing N-best lists in order to increase their potential of containing better translations. New entries are generated by means of a word-based n-gram language model estimated on the N-best entries. Experimental results on the NIST Chinese-to-English task show that better N-best lists can be obtained which also result in systematic BLEU score improvements in the re-scoring step.

1. Introduction

In Statistical Machine Translation (SMT), performance improvements are often reported by applying two processing steps (Federico and Bertoldi, 2005; Koehn et al., 2003). In the first step, a decoding algorithms is applied that generates an N-best list of translation hypotheses. In the second step, the final translation is computed by re-ranking the N-best translations through additional scores, computed with more sophisticated feature functions. Clearly, a fundamental assumption of the two step approach is that the generated N-best list contains better translations than the best one found by the decoder. The aim of the additional feature functions is indeed to reward better translations found among the N-best entries of the decoder.

The reason for applying two steps instead of one is that not all available feature functions can be efficiently implemented into the decoder. In fact, not all of them can be decomposed into local scores that can be computed on partial translation hypotheses. Moreover, recently feature functions have been proposed that are estimated directly on the N-best list. In particular, (Chen et al., 2005; Zens and Ney, 2006) have recently reported performance improvements by computing posterior probabilities through n-gram language models (LMs) estimated on the N-best translations.

This paper proposes an intermediate step in the chain. Before applying re-scoring, the N-best list is further expanded by applying a generative statistical n-gram LM, estimated on the N-best list itself. In particular, the LM is used to generate M new and different target strings, that do not occur in the N-best list.

We applied this technique to a well performing baseline for Chinese-to-English translation, trained under the large-data condition set by the NIST MT Workshop. Experiments were carried out to test whether the N+M-best lists generated by our method are better than those the decoder would generate, and to verify if the second decoding step achieves better MT performance by exploiting the expanded list.

The remaining part of this paper is organized as follow. Section 2 presents the phrase-based SMT system we work with. Section 3 introduces the new hypotheses generation algorithm. Section 4 describes experiments and analyzes results. A discussion and conclusions end the paper.

2. SMT Process

Given a string $\mathbf{f}$ in the source language, the goal of SMT is to select the string $\mathbf{e}$ in the target language which maximizes the posterior distribution $\Pr(\mathbf{e} \mid \mathbf{f})$. In phrase-based translation, words are no longer the only units of translation, but they are complemented by strings of consecutive words, the phrases. By assuming a log-linear model (Berger et al., 1996; Och and Ney, 2002) and by introducing the concept of word alignment (Brown et al., 1993), the optimal translation can be searched for with the criterion:

$$\tilde{\mathbf{e}}^* = \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a}} \sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a}),$$

where $\tilde{\mathbf{e}}$ represents a string of phrases in the target language, $\mathbf{a}$ an alignment from the words in $\mathbf{f}$ to the phrases in $\tilde{\mathbf{e}}$, and $h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})$ $r = 1, \ldots, R$ are *feature functions*, designed to model different aspects of the translation process. We performed the "argmax" operation of the above equation by means of the decoder available in `Moses`,[1] an open source toolkit for SMT. Besides the decoder, `Moses` provides tools for training translation and lexicalized reordering models, and a minimum error training procedure for estimating optimal interpolation weights.

Actually, the decoder is used to generate not only the best translation of the given source sentence but the list of N-best translation hypotheses. The list is then augmented through the procedure detailed in the following section and finally re-ranked. Re-ranking is computed by means of a new log-linear model that includes new feature functions which are listed in Section 4. Finally, the resulting best scoring entry is output as final translation.

3. N-best List Expansion

The rationale behind our approach is that alternative translation hypotheses can be obtained by combining substrings

[1] http://www.statmt.org/moses/

occurring in the N-best list. In fact, it could be the case that the best scoring translation in the list is wrong and that a correct translation could be obtained by replacing some of its words with portions taken from other translations in the N-best lists.

The actual implementation of our idea implies the enlargement of the N-best list with new hypotheses generated through an n-gram LM estimated on the N-best list itself. Assuming that a partial hypothesis (e_1 e_2 e_3 e_4 e_5) is available (see Figure 1), this can be expanded by one word through an n-gram whose first $n-1$ words match the last $n-1$ words of the hypothesis.

partial hyp	e_1	e_2	e_3	e_4	e_5	
n-gram				e_4	e_5	e_6
new partial hyp	e_1	e_2	e_3	e_4	e_5	e_6

Figure 1: Expansion of a partial hypothesis via a matching n-gram.

```
Input: N-best translation hypotheses of some source string

Initialization:
    $n$ = some integer        (size of $n$-grams)
    P_HYP = $\emptyset$              (set of partial translations)
    F_HYP = $\emptyset$              (set of complete translations)
    NGRAM = $\emptyset$              (set of $n$-grams in N-best list)

 1  for each translation hypothesis e $\in$ N-best list
 2      do
 3          insert starting $n$-gram of e in P_HYP
 4          insert each $n$-gram of e in NGRAM
 5  while not empty P_HYP
 6      do
 7          extract one e from P_HYP
 8          compute EXP $\subseteq$ NGRAM such that
 9              $\forall n$-gram $\in$ EXP: match(e, $n$-gram)
10          $\forall$ $n$-gram $\in$ EXP do
11              e$'$ = expand(e, $n$-gram)
12              if not too long e$'$
13                  if complete e$'$
14                      if not too short e$'$
15                          F_HYP = F_HYP $\cup$ {e$'$}
16                  else P_HYP = P_HYP $\cup$ {e$'$}
17  return F_HYP
```

Figure 2: N-best list expansion algorithm.

This mechanism is the core of the expansion algorithm presented in Figure 2. It is applied to the N-best list of translation hypotheses of a given source sentence. First, all n-grams occurring in the list are collected (line 4). Moreover, the set of partial hypotheses is initialized with the first n-gram of each entry of the list (line 3).

The expansion of a partial hypothesis (line 7) starts by computing the set of n-grams matching its last $n-1$ words (lines 8,9). Then, the partial translation is expanded by appending the last words of each of these n-grams (lines 10,11). If expanded hypotheses are not too long with respect to the source sentence (line 12), it is checked if they are complete translations or not (line 13). In the first case, if they are not too short (line 14) they are added to the set of translations to be output (line 15). If they are still partial, they are added to the proper set (line 16). Notice, that a hypothesis is final if it ends with the special end-of-sentence symbol that occurs at the end of all N-best strings.

Eventually, the set of partial translation becomes empty and complete translations are output (line 17).

The algorithm prunes unlikely hypotheses on the basis of two length thresholds (lines 12 and 14). Their value derives from training data statistics: in our case, for example, we observed that more than 90% of Chinese and English sentence pairs has a source to target length ratio falling in the 0.9-1.5 interval. Of course, the value of both thresholds must be estimated for the language-pair at hand.

In fact, another pruning method is applied that is not explicitly mentioned in Figure 2. To avoid the case of a too large number of expansions, pruning of partial strings is applied on the basis of their likelihood. In particular, hypotheses are scored through the log-linear combination of two models. The first model takes into account the frequency of the contained n-grams, starting from 1-grams up to 4-grams, according to the statistics gathered from the N-best list (Chen et al., 2005). The second model computes the n-gram posterior probability as proposed by (Zens and Ney, 2006). Actually, other meaningful feature functions could be interpolated to score hypotheses such as IBM models, language models, etc. Nevertheless, we did not observe any significant benefit in adding further models in our experiments.

Finally, complete translations are pruned in order to not exceed the fixed amount of M. Moreover, a just generated complete translation is added to the set (line 15) only if it is distinct from any translation of the original N-best list. In this way, the expansion algorithm only generates new translation hypotheses.

When generation of new hypotheses is finished, they are joined to the original set of translation hypotheses so that an enlarged N+M-best translations list is built. Then, rescoring and re-ranking are applied by using additional feature functions and the top ranking candidate is selected as the final translation.

4. Experiments

4.1. Task

The task chosen for our experiments is the translation of news from Chinese to English, as proposed by the NIST MT Evaluation Workshops.[2] A translation baseline system was trained according to the *large-data* condition. In particular, all the allowed bilingual corpora have been used for estimating the phrase-table and a lexicalized reordering model. The target side of these texts was also employed for the estimation of a 5-gram LM, henceforth named "large", smoothed via the improved Kneser-Ney method (Chen and Goodman, 1999). An additional, much larger, 5-gram LM was instead trained on the so-called English Gigaword corpus, one of the allowed monolingual resources for this task ("giga" LM).

[2] www.nist.gov/speech/tests/mt/

set	type	|W|	
		source	target
large	parallel	83.1M	87.6M
giga	monolingual	-	1.76G
NIST 02	dev	23.7K	26.4K
NIST 03	test	25.6K	28.5K
NIST 04	test	51.0K	58.9K
NIST 05	test	31.2K	34.6K

Table 1: Statistics of training, dev. and test sets. Evaluation sets of NIST campaigns include 4 references: in table, average lenghts are provided.

LM	1-gram	2-gram	3-gram	4-gram	5-gram
dev	17K	153K	71K	64K	-
large	0.3M	5.3M	4.8M	6.3M	6.1M
giga	4.5M	64.4M	127.5M	228.8M	288.6M

Table 2: Statistics of LMs.

Automatic translation was performed by means of Moses which, among other things, permits the contemporary use of more LMs, feature we exploited in our experiments as specified later.

Optimal interpolation weights for the log-linear model were estimated by running the minimum error training algorithm (Och, 2004) available in the Moses toolkit, on the evaluation set of the NIST 2002 campaign. Tests were performed on the test sets of the 2003, 2004, and 2005 NIST evaluations .

Table 1 gives figures about training, development and test corpora, while Table 2 provides main statistics of the LMs employed during decoding: besides “large” and “giga” LMs, a third 4-gram LM was trained on the English side of the development sets (“dev”).

MT performance are provided in terms of case-insensitive BLEU and NIST scores, as computed with the NIST scoring tool.

4.2. Setup

For the generation of N-best translation lists, we run Moses with the maximum reordering distance set to 6 and the following feature functions:

- phrase translation model, with phrases including at most 7 words
- 3 LMs, namely “dev”, “large” and “giga”
- lexicalized distortion model, trained specifying the option “orientation-bidirectional-fe” (Koehn et al., 2005)
- word and phrase penalties, for balancing the length of translations with that of input sentences.

Once N-best lists are available, they are expanded through the algorithm described in Section 3 by setting $n = 4$. Then, original (1) and updated (2) lists are re-scored and then re-ranked by applying some of the following feature functions, as specified in brackets:

list	BLEU%	NIST	PER%	WER%
100-best	36.64	9.571	38.79	59.53
50+50-best	37.53	9.678	37.76	57.93

Table 4: Oracle scores on 100-best lists, with and w/o n-gram expansion, on test set NIST 2005.

- decoding feature functions (applied to lists of type 1)
- average of N-best translation scores output by the decoder (applied to lists of type 2)
- direct and inverse IBM model 1 and 3 lexicons, over all possible alignments (1, 2)
- frequency of its n-grams (n=1,2,3,4) within the N-best translations (1, 2)
- frequency of its n-grams (n=1,2,3,4) within the N+M-best translations (2)
- “dev” and “large” LMs (2)
- n-gram posterior probabilities within the original N-best translations (Zens and Ney, 2006) (1, 2)
- n-gram posterior probabilities within the expanded N+M-best translations (2)
- sentence length posterior probabilities within the original N-best translations (Zens and Ney, 2006) (1, 2)
- sentence length posterior probabilities within the expanded N+M-best translations (2).

Re-ranking of translation lists is performed by a log-linear model with interpolation weights estimated again on the NIST 02 evaluation set. Weight estimation is carried out by optimizing a combination of BLEU and NIST scores (Chen et al., 2005) through an iterative combination of the Simplex (Press et al., 2002) and Powell (Powell, 1964) algorithms. It is worth underlining that different interpolation weights were estimated for re-scoring the original N-best lists and the augmented N+M-entry lists. Moreover, the number of feature functions used in the former log-linear model is definitely larger than for the model re-scoring N+M-best lists. For instance, the model re-scoring N-best lists embeds all single feature functions used by the decoder, while the model re-scoring the expanded lists concentrates the decoder’s feature functions into one single global statistics. Moreover, for the additional M translations generated by the LM, the decoder score is replaced by a mean score computed over the N-best translations. The lack of detailed feature functions from the decoder is however compensated, to some extent, by the use of target LM feature functions, exactly as in the decoder.

4.3. Results

Re-scoring experiments were carried out with three different lists of translations: (i) 5K-best translations computed by Moses; (ii) 3K-best translations from Moses plus 2K from expansion; and (iii) 5K from Moses plus 5K from expansion. In all the settings, lists do not contain duplicates.

Translation performance are reported in Table 3. The columns with header “1-best” report scores of the translations output by the decoder; the others columns show in-

evaluation	1-best		5K-best		3K+2K-best		5K+5K-best	
set	BLEU%	NIST	BLEU%	NIST	BLEU%	NIST	BLEU%	NIST
NIST 02 (dev)	35.06	9.564	35.43	9.705	35.86	9.688	35.98	9.693
NIST 03	33.62	9.270	34.01	9.396	34.48	9.399	34.60	9.385
NIST 04	35.04	9.746	35.34	9.821	35.66	9.767	35.78	9.763
NIST 05	31.92	9.005	32.23	9.114	32.54	9.091	32.68	9.081

Table 3: Translation results for different NIST evaluation sets.

stead results obtained by re-scoring different lists of translation hypotheses.

First of all, it can be noted that both NIST and BLEU scores of the first decoding step are improved by about 1% relative through re-scoring of 5K-best translations. If re-scoring is instead applied to 5K entries obtained by expanding 3K-best translations with 2K translations, a further relative improvement of BLEU by about 1% is observed. By expanding 5K-best translations with other 5K translations only limited gains of the BLEU score are observed. On the other hand, the re-scoring of expanded lists does not yield any further increment of NIST scores.

In order to better assess the quality of the new generated hypotheses, we analyzed in detail the translations of the NIST 2005 evaluation set. After re-ranking 3K+2K entries, it resulted that 15% (160 out of 1082) of best scored outputs were generated by n-gram expansion, showing that new generated translations are quite often the re-scoring winner. From another viewpoint, Table 4 reports oracle scores computed on two 100-best lists: (i) containing only translations from decoder; (ii) containing 50 translations from decoder plus 50 translations generated by expansion. The oracle chooses in both cases the translation with the lowest word error rate with respect to the references. It is worth to noticing that N-best expansion improves scores of BLEU, PER and WER by about 2.5% relative, and of NIST by 1% relative, showing that even if new generated translations are not the re-scoring winner, they are good translations in comparison of original ones.

Some examples of translations obtained under all compared conditions are shown in Table 5. As a confirmation of the versatility of the n-gram expansion mechanism, it can be noted that the generation of new hypotheses can involve all possible operations:

substitutions: in the first example, the word “say” is substituted by “described”; in the third example, the sequence “to 10.1 percent from 10.3 percent” is substituted by “from 10.1 percent to 10.3 % “.

insertions: in the second example, the word “china” and the sequence “animal and plant” are inserted.

deletion: in the fifth example, the word “level” is deleted.

re-ordering: in the forth example, the word “them” is re-ordered.

5. Discussion

The method we propose is innovative; to the best of our knowledge, no previous work presented a similar approach. The N-best expansion step exploits the work done by the decoder, which explores a huge search space but is guided by a set of constraints and scores computed through suitable feature functions. Constraints used by the decoder mainly limit the possible word re-ordering and the translation alternatives for single words and phrases.

On the contrary, our expansion step of the N-best translations (almost) fully exploits the search space of target strings, that can be generated by an n-gram LM. As a result, in principle it can generate translation hypotheses outside the scope of the decoder’s constraints. For instance, it is easy to verify that a low-order LM (e.g. a bigram LM) permits long word movements and the creation of phrases which are not contained in the phrase-table.

The algorithm we have proposed for expanding N-best lists is simple and intuitive, but nevertheless effective in improving the quality of a two-pass SMT system. In fact, small but consistent improvements were measured on various evaluation sets on the challenging Chinese-to-English NIST MT task over a state-of-the-art performing baseline.

We are aware of the criticisms about the BLEU score expressed in (Callison-Burch et al., 2006), which could particularly apply to our technique. In their paper, Burch *et. al* showed that for a translation output there are many possible variants, based on word permutations, that would each receive a similar BLEU score. Some variations could even correspond to higher scores but not to any genuine improvement in translation quality. The same could indeed apply to the translations computed by the expansion step, which in practice generates new word arrangements from the N-best list. We cannot reject this claim, as we did not manually inspect all the translations. However, from one hand our oracle experiments (Table 4) show that the expansion step produces N-best lists with lower word-error rates, a score which is very sensitive to the word order; on the other hand, our BLEU improvements are small but consistent over four test sets (Table 3), which considerably reduces the chance that our improvements are random.

We think nevertheless that our approach deserves further investigation. As a next step, we will check if the generation of new translations by n-gram expansion is effective with other language pairs and tasks as well.

Some weaknesses of the procedure will be faced, too. In particular, we plan to improve the log-linear model employed to score partial hypotheses during the expansion step. Alternative feature functions will be considered as well as a procedure for estimating the weights of the log-linear model.

Additional feature functions for the final re-scoring stage will be investigated as well. In particular, word alignment information for the new translations will be computed on

1	1-best	election observers say that this is a free and fair election . ”
	+expansion	election observers described this is a free and fair election . ”
	reference	observers of the election described it as a largely free and fair election . ”
2	1-best	(international) and chile signed a memorandum on the implementation of health measures
	+expansion	(international) china and chile signed a memorandum on the implementation of animal and plant health measures
	reference	(international) china and chile sign memorandum on the implementation of animal and plant sanitation measures
3	1-best	the seasonally adjusted unemployment rate rose to 10.1 percent from 10.3 percent .
	+expansion	meanwhile , the seasonally adjusted unemployment rate increased from 10.1 percent to 10.3 % .
	reference	meanwhile , seasonally adjusted unemployment rate rose to 10.3 % from 10.1 % .
4	1-best	the turkish judiciary has the right to arrest them , sentences and imprisonment .
	+expansion	the turkish judiciary has the right to arrest, sentence and imprisonment them .
	reference	the turkish judiciary departments have a right to arrest , sentence and imprison them .
5	1-best	it is estimated that the euro zone economic growth rate this year will reach 2 % to 2.5 % level .
	+expansion	it is estimated that the euro zone economic growth rate this year will reach 2 % to 2.5 % .
	reference	it is estimated that the economic growth rate in the european zone will reach 2 % to 2.5 % .

Table 5: Translations output by the decoder and after re-scoring expanded N-best lists.

the basis of translation models, e.g. the competitive linking algorithm (Melamed, 2000). In this way, alternative word re-ordering models (Chen et al., 2006) could be applied to enrich the global score.

Finally, a deeper investigation will be performed to see whether the better translations, coming out from the final re-scoring step, were or not inside the search space explored by the phase-based decoder.

6. Conclusions

We have presented a novel method for improving the quality of N-best translations computed by a state-of-the-art phrase-based decoder. New translations are added to the N-best list by means of a generative LM trained on the N-best list. In this way, alternative translations can be obtained which contain word re-orderings and phrase structures not considered by the search algorithm. Mild constraints have been introduced to discard too short, too long, or too unlikely translations. Experiments carried out on the NIST Chinese-to-English task show that the additional translations reduce the word error rate of the full set of hypotheses more than the decoder is able to do. Moreover, re-scoring the augmented set of translations results in consistently improved scores by an already well-performing baseline.

7. Acknowledgements

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

8. References

A. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–312.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of Bleu in machine translation research. In *Proceedings of EACL*, pages 249–256, Trento, Italy.

S. F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13):359–393.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. In *Proceedings of IWSLT*, pages 98–104, Pittsburgh, PA, USA.

B. Chen, M. Cettolo, and M. Federico. 2006. Reordering Rules for Phrase-based Statistical Machine Translation. In *Proceedings of IWSLT*, pages 182–189, Kyoto, Japan, Nov.

M. Federico and N. Bertoldi. 2005. A word-to-phrase statistical translation model. *ACM Transaction on Speech Language Processing*, 2(2):1–24.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL 2003*, pages 127–133, Edmonton, Canada.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of IWSLT*, Pittsburgh, PA, USA.

I. Dan Melamed. 2000. Models of Translational Equivalence among Words. *Computational Linguistics*, 26(2):221–249.

F. J. Och and H. 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, PA, Philadelphia, USA, July.

F. J. Och. 2004. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, Sapporo, Japan.

M. J. D. Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162.

W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. 2002. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press.

R. Zens and H. Ney. 2006. N-Gram Posterior Probabilities for Statistical Machine Translation. In *Proceedings of the NAACL Workshop on SMT*, New York, USA.