

Analysis and System Combination of Phrase- and N -gram-based Statistical Machine Translation Systems

Marta R. Costa-jussà¹, Josep M. Crego^{1,2}, David Vilar²
José A. R. Fonollosa¹, José B. Mariño¹ and Hermann Ney²

¹TALP Research Center (UPC), Barcelona 08034, Spain
{mruiz, jmcrego, adrian, canton}@gps.tsc.upc.edu

²RWTH Aachen University, Aachen D-52056, Germany
{vilar, ney}@i6.informatik.rwth-aachen.de

Abstract

In the framework of the TC-STAR project, we analyze and propose a combination of two Statistical Machine Translation systems: a phrase-based and an N -gram-based one. The exhaustive analysis includes a comparison of the translation models in terms of efficiency (number of translation units used in the search and computational time) and an examination of the errors in each system's output. Additionally, we combine both systems, showing accuracy improvements.

1 Introduction

Statistical machine translation (SMT) has evolved from the initial word-based translation models to more advanced models that take the context surrounding the words into account. The so-called phrase-based and N -gram-based models are two examples of these approaches (Zens and Ney, 2004; Mariño et al., 2006).

In current state-of-the-art SMT systems, the phrase-based or the N -gram-based models are usually the main features in a log-linear framework, reminiscent of the maximum entropy modeling approach.

Two basic issues differentiate the N -gram-based system from the phrase-based one: the training data is sequentially segmented into bilingual units; and the probability of these units is estimated as a bilingual N -gram language model. In the phrase-based model, no monotonicity restriction is imposed on the segmentation and the probabilities are normally estimated simply by relative frequencies.

This paper extends the analysis of both systems performed in (Crego et al., 2005a) by additionally performing a manual error analysis of both systems, which were the ones used by UPC and RWTH in the last TC-STAR evaluation.

Furthermore, we will propose a way to combine both systems in order to improve the quality of translations.

Experiments combining several kinds of MT systems have been presented in (Matusov et al., 2006), based only on the single best output of each system. Recently, a more straightforward approach of both systems has been performed in (Costa-jussà et al., 2006) which simply selects, for each sentence, one of the provided hypotheses.

This paper is organized as follows. In section 2, we briefly describe the phrase and the N -gram-based baseline systems. In the next section we present the evaluation framework. In Section 4 we report a structural comparison performed for both systems and, afterwards, in Section 5, we analyze the errors of both systems. Finally, in the last two sections we rescore and combine both systems, and the obtained results are discussed.

2 Baseline Systems

2.1 Phrase-based System

The basic idea of phrase-based translation is to segment the given source sentence into units (here called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

In order to train these phrase-based models, an alignment between the source and target training sentences is found by using the standard IBM models in both directions (source-to-target and target-to-source) and combining the two obtained alignments. Given this alignment an extraction of contiguous phrases is carried out, specifically we extract all phrases that fulfill the following restrictions: all source (target) words within the phrase are aligned only to target (source) words within the phrase.

The probability of these phrases is normally estimated by relative frequencies, normally in both directions, which are then combined in a log-linear way.

2.2 N -gram-based System

In contrast with standard phrase-based approaches, the N -gram translation model uses *tuples* as bilingual units whose probabilities are estimated as an N -gram language model (Mariño et al., 2006). This model approximates the joint probability between the source and target languages by using N -grams.

Given a word alignment, tuples define a unique and monotonic segmentation of each bilingual sentence, building up a much smaller set of units than with phrases and allowing N -gram estimation to account for the history of the translation process (Mariño et al., 2006).

2.3 Feature functions

Both baseline systems are combined in a log-linear way with several additional feature functions: a target language model, a forward and a backward lexicon model and a word bonus are common features for both systems. The phrase-based system also introduces a phrase bonus model.

3 Evaluation framework

The translation models presented so far were the ones used by UPC and RWTH in the second evaluation campaign of the TC-STAR project. The goal of this project is to build a speech-to-speech translation system that can deal with real life data.

The corpus consists of the official version of the speeches held in the European Parliament Plenary Sessions (EPPS), as available on the web page of the European Parliament. Table 1 shows some statistics.

The following tools have been used for building both systems: Word alignments were computed using GIZA++ (Och, 2003), language models were estimated using the SRILM toolkit (Stolcke, 2002), decoding was carried out by the free available MARIE decoder (Crego et al., 2005b) and the optimization was performed through an in-house implementation of the simplex method (Nelder and Mead, 1965).

		Spanish	English
Train	Sentences	1.2M	
	Words	32M	31M
	Vocabulary	159K	111K
Dev	Sentences	1 122	699
	Words	26K	21K
Test	Sentences	1 117	894
	Words	26K	26K

Table 1: Statistics of the EPPS Corpora.

4 Structural comparison

Both approaches aim at improving accuracy by including word context in the model. However, the implementation of the models are quite different and may produce variations in several aspects.

Table 2 shows the effect on decoding time introduced through different settings of the beam size. Additionally, the number of available translation units is shown, corresponding to number of available phrases for the phrase-based system and 1gram, 2gram and 3gram entries for the N -gram-based system. Results are computed on the development set.

Task	Beam	Time(s)	Units
es→en	50	2,677	537k
	10	852	
	5	311	
en→es	50	2,689	594k
	10	903	
	5	329	
es→en	50	1,264	104k 288k 145k
	10	281	
	5	138	
en→es	50	1,508	118k 355k 178k
	10	302	
	5	155	

Table 2: Impact on efficiency of the beam size in PB (top) and NB system (bottom).

As it can be seen, the number of translation units is similar in both tasks for both systems ($537k \sim 537k$ for Spanish to English and $594k \sim 651k$ for English to Spanish) while the time consumed in decoding is clearly higher for the phrase-based system. This can be explained by the fact that in the phrase-based approach, the same translation can be hypothesized following several segmentations of the input sentence, as phrases appear (and are collected) from multiple segmentations of the training sentence pairs. In other words, the search graph seems to be overpopulated under the phrase-based approach.

Table 3 shows the effect on translation accuracy regarding the size of the beam in the search. Results are computed on the test set for the phrase-based and N -gram-based systems.

Results of the N -gram-based system show that decreasing the beam size produces a clear reduction of the accuracy results. The phrase-based system shows that accuracy results remain very similar under the different settings. The reason is found on how translation models are used in the search. In the phrase-based approach, every partial hypothesis

Task	Beam	BLEU	NIST	mWER
es→en	50	51.90	10.53	37.54
	10	51.93	10.54	37.49
	5	51.87	10.55	37.47
en→es	50	47.75	9.94	41.20
	10	47.77	9.96	41.09
	5	47.86	10.00	40.74
es→en	50	51.63	10.46	37.88
	10	51.50	10.45	37.83
	5	51.39	10.45	37.85
en→es	50	47.73	10.08	40.50
	10	46.82	9.97	41.04
	5	45.59	9.83	41.04

Table 3: Impact on accuracy of the beam size in PB (top) and NB system (bottom).

is scored uncontextualized, hence, a single score is used for a given partial hypothesis (phrase). In the N -gram-based approach, the model is intrinsically contextualized, which means that each partial hypothesis (tuple) depends on the preceding sequence of tuples. Thus, if a bad sequence of tuples (bad scored) is composed of a good initial sequence (well scored), it is placed on top of the first stacks (beam) and may cause the pruning of the rest of hypotheses.

5 Error analysis

In order to better assess the quality and the differences between the two systems, a human error analysis was carried out. The guidelines for this error analysis can be found in (Vilar et al., 2006). We randomly selected 100 sentences, which were evaluated by bilingual judges.

This analysis reveals that both systems produce the same kind of errors in general. However some differences were identified. For the English to Spanish direction the greatest problem is the correct generation of the right tense for verbs, with around 20% of all translation errors being of this kind. Reordering also poses an important problem for both phrase and N -gram-based systems, with 18% or 15% (respectively) of the errors falling into this category. Missing words is also an important problem. However, most of them (approximately two thirds for both systems) are filler words (i.e. words which do not convey meaning), that is, the meaning of the sentence is preserved. The most remarkable difference when comparing both systems is that the N -gram based system produces a relatively large amount of extra words (approximately 10%), while for the phrase-based system, this is only a minor problem (2% of the errors). In contrast the phrase-based system has

more problems with incorrect translations, that is words for which a human can find a correspondence in the source text, but the translation is incorrect.

Similar conclusions can be drawn for the inverse direction. The verb generating problem is not so acute in this translation direction due to the much simplified morphology of English. An important problem is the generation of the right preposition.

The N -gram based system seems to be able to produce more accurate translations (reflected by a lower percentage of translation errors). However, it generates too many additional (and incorrect words) in the process. The phrase-based system, in contrast, counteracts this effect by producing a more direct correspondence with the words present in the source sentence at the cost of sometimes not being able to find the exact translation.

6 System Rescoring and Combination

Integration of both output translations in the search procedure is a complex task. Translation units of both models are quite different and generation histories pose severe implementation difficulties. We propose a method for combining the two systems at the level of N -best lists.

Some features that are useful for SMT are too complex for including them directly in the search process. A clear example are the features that require the entire target sentence to be evaluated, as this is not compatible with the pruning and recombination procedures that are necessary for keeping the target sentence generation process manageable. A possible solution for this problem is to apply sentence level re-ranking by using N -best lists.

6.1 Rescoring Criteria

The aim of the rescoring procedure is to choose the best translation candidate out of a given set of N possible translations. In our approach this translation candidates are produced independently by both of the systems and then combined by a simple concatenation¹. In order for the hypothesis to have a comparable set of scores, we perform an additional “cross-rescoring” of the lists.

Given an N -best list of the phrase-based (N -gram-based) system, we compute the cost of each target sentence of this N -best list for the N -gram-based (phrase-based) system. However this computation is not possible in all cases. Table 4 shows the percentage of target sentences that the N -gram-based

¹With removal of duplicates.

(phrase-based) system is able to produce given an N -best list of target sentences computed by the phrase-based (N -gram-based) system. This percentage is calculated on the development set.

The vocabulary of phrases is bigger than the vocabulary of tuples, due to the fact that phrases are extracted from multiple segmentations of the training sentence pairs. Hence, the number of sentences reproduced by the N -gram-based system is smaller than the number of sentences reproduced by the phrase-based system. Whenever a sentence can not be reproduced by a given system, the cost of the worst sentence in the N -best list is assigned to it.

Task	N -best	% NB	% PB
es→en	1000	37.5	57.5
en→es	1000	37.2	48.6

Table 4: Sentences (%) produced by each system.

6.2 Results

Table 5 shows results of the rescoring and system combination experiments on the test set. The first two rows include results of systems non-rescored and PB (NB) rescored by NB (PB). The third row corresponds to the system combination. Here, PB (NB) rescored by NB (PB) are simply merged and ranked by rescored score.

System	N -best	BLEU	NIST	mWER
Spanish-to-English				
PB	1	51.90	10.54	37.50
PB	1000	52.55	10.61	37.12
NB	1	51.63	10.46	37.88
NB	1000	52.25	10.55	37.43
PB+NB	2	51.77	10.49	37.68
PB+NB	2000	52.31	10.56	37.32
English-to-Spanish				
PB	1	47.75	9.94	41.2
PB	1000	48.46	10.13	39.98
NB	1	47.73	10.09	40.50
NB	1000	48.33	10.15	40.13
PB+NB	2	48.26	10.05	40.61
PB+NB	2000	48.54	10.16	40.00

Table 5: Rescoring and system combination results.

7 Discussion

The structural comparison has shown on the one hand that the N -gram-based system outperforms the phrase-based in terms of search time efficiency by avoiding the overpopulation problem presented

in the phrase-based approach. On the other hand the phrase-based system shows a better performance when decoding under a highly constrained search.

A detailed error analysis has also been carried out in order to better determine the differences in performance of both systems. The N -gram based system produced more accurate translations, but also a larger amount of extra (incorrect) words when compare to the phrase-based translation system.

In section 6 we have presented a system combination method using a rescoring feature for each SMT system, i.e. the N -gram-based feature for the phrase-based system and vice-versa. For both systems, considering the feature of the opposite system leads to an improvement of BLEU score.

References

- M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov J.A.R. Fonollosa, J.B. Mariño, and R. Banchs. 2006. Talp phrase-based statistical machine translation and talp system combination the iwslt 2006. *IWSLT06*.
- J. M. Crego, M. R. Costa-jussà, J. Mariño, and J. A. Fonollosa. 2005a. N-gram-based versus phrase-based statistical machine translation. *IWSLT05*, October.
- J.M. Crego, J. Mariño, and A. de Gispert. 2005b. An Ngram-based statistical machine translation decoder. *ICSLP05*, April.
- J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549.
- E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. *EACL06*, pages 33–40.
- J.A. Nelder and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.
- F.J. Och. 2003. Giza++ software. <http://www-i6.informatik.rwth-aachen.de/~och/software/giza++.html>.
- A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.
- David Vilar, Jia Xu, Luis Fernando D’Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *LREC06*, pages 697–702, Genoa, Italy, May.
- Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *HLT04*, pages 257–264, Boston, MA, May.