LIMSI's statistical translation systems for WMT'09

Alexandre Allauzen, Josep Crego, Aurélien Max and François Yvon
LIMSI/CNRS and Université Paris-Sud 11, France
BP 133, 91403 Orsay Cédex
firstname.lastname@limsi.fr

Abstract

This paper describes our Statistical Machine Translation systems for the WMT09 (en:fr) shared task. For this evaluation, we have developed four systems, using two different MT Toolkits: our primary submission, in both directions, is based on Moses, boosted with contextual information on phrases, and is contrasted with a conventional Moses-based system. Additional contrasts are based on the Ncode toolkit, one of which uses (part of) the English/French GigaWord parallel corpus.

1 Introduction

This paper describes our Statistical Machine Translation systems for the WMT09 (en:fr) shared task. For this evaluation, we have developed four systems, using two different MT toolkits: our primary submission, in both direction, is based on Moses, boosted with contextual information on phrases; we also provided a contrast with a vanilla Moses-based system. Additional contrasts are based on the N-code decoder, one of which takes advantage of (part of) the English/French GigaWord parallel corpus.

2 System architecture and resources

In this section, we describe the main characteristics of the baseline phrase-based systems used in this evaluation and the resources that were used to train our models.

2.1 Pre- and post-processing tools

All the available textual corpora were processed and normalized using in-house text processing tools. Our last year experiments (Déchelotte et al., 2008) revealed that using better normalization tools provides a significant reward in BLEU, a fact that we could observe again this year. The downside is the need to post-process our outputs so as to "detokenize" them for scoring purposes, which is unfortunately an error-prone process.

Based again on last year's experiments, our systems are built in "true case": the first letter of each sentence is lowercased when it should be, and the remaining tokens are left as is.

Finally, the N-code (see 2.5) and the context-aware (see 3) systems require the source to be morpho-syntactically analysed. This was performed using the TreeTagger[1] for both languages.

2.2 Alignment and translation models

Our baseline translation models (see 2.4 and 2.5) use all the parallel corpora distributed for this evaluation: Europarl V4, news commentary (2006-2009) and the additional news data, totalling 1.5M sentences. Our preliminary attempts with larger translation models using the GigaWord corpus are reported in section 3.2. All these corpora were aligned with GIZA++[2] using default settings.

2.3 Language Models

To train our language models (LMs), we took advantage of the *a priori* information that the test set would be of newspaper/newswire genre. We

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

[2] http://www.fjoch.com/GIZA++.html.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 100–104,
Athens, Greece, 30 March – 31 March 2009.

	Source	Period	M. words
En	News texts	1994-06	3 317
	BN transcripts	2000-07	341
	WMT		86
Fr	Newswires	1994-07	723
	Newspapers	1987-06	486
	WEB	2008	23
	WMT		46
	News-train08		167

Table 1: Corpora used to train the target language models in English and French.

thus built much larger LMs for translating both to French and to English, and optimized their combination on the first part of the official development data (dev2009a).

Corpora and vocabulary Statistics regarding the training material are summarized in table 1 in terms of source, time period, and millions of occurrences. "WMT" stands for all text provided for the evaluation. Development sets and the large training corpora (news-train08 and the GigaWord corpus) were not included. Altogether, these data contain a total number of 3.7 billion tokens for English and 1.4 billion tokens for French.

To estimate such large LMs, a vocabulary was first defined for both languages by including all tokens in the WMT parallel data. This initial vocabulary of 130K words was then extended by adding the most frequent words observed in the additional training data. This procedure yielded a vocabulary of one million words in both languages.

Language model training The training data were divided into several sets based on dates on genres (resp. 7 and 9 sets for English and French). On each set, a standard 4-gram LM was estimated from the 1M word vocabulary with in-house tools using absolute discounting interpolated with lower order models. The resulting LMs were then linearly interpolated using interpolation coefficients chosen so as to minimise perplexity of the development set (dev2009a). Due to memory limitations, the final LMs were pruned using perplexity as pruning criterion.

Out of vocabulary word and perplexity To evaluate our vocabulary and LMs, we used the official devtest and test sets. The out-of-vocabulary (OOV) rate was drastically reduced by increasing the vocabulary size, the mean OOV rate decreasing from 2.5% to 0.7%, a trend observed in both languages.

For French, using a small LM trained on the "WMT" data only resulted in a perplexity of 301 on the devtest corpus and 299 on the test set. Using all additional data yielded a large decrease in perplexity (106 on the devtest and 108 on the test); again the same trend was observed for English.

2.4 A Moses baseline

Our baseline system was a vanilla phrase-based system built with Moses (Koehn et al., 2007) using default settings. Phrases were extracted using the 'grow-diag-final-and' heuristics, using a maximum phrase length of 7; non-contextual phrase scores contain the 4 translation model scores, plus a fixed phrase penalty; 6 additional scores parameterize the lexicalized reordering model. Default decoding options were used (20 alternatives per phrase, maximum distortion distance of 7, etc.)

2.5 A N-code baseline

N-code implements the *n*-gram-based approach to Statistical Machine Translation (Mariño et al., 2006). In a nutshell, the translation model is implemented as a stochastic finite-state transducer trained using a *n*-gram model of (source,target) pairs (Casacuberta and Vidal, 2004). Training such a model requires to reorder source sentences so as to match the target word order. This is also performed via a stochastic finite-state reordering model, which uses part-of-speech information to generalise reordering patterns beyond lexical regularities. The reordering model is trained on a version of the parallel corpora where the source sentences have been reordered via the unfold heuristics (Crego and Mariño, 2007). A conventional n-gram language model of the target language provides the third component of the system.

In all our experiments, we used 4-gram reordering models and bilingual tuple models built using Kneser-Ney backoff (Chen and Goodman, 1996). The maximum tuple size was also set to 7.

2.6 Tuning procedure

The Moses-based systems were tuned using the implementation of minimum error rate training (MERT) (Och, 2003) distributed with the Moses decoder, using the development corpus (dev2009a). For the context-less systems, tuning concerned the 14 usual weights; tuning the

22 weights of the context-aware systems (see 3.1) proved to be much more challenging, and the weights used in our submissions are probably far from optimal. The N-code systems only rely on 9 weights, since they dispense with the lexical reordering model; these weights were tuned on the same dataset, using an in-house implementation of the simplex algorithm.

3 Extensions

3.1 A context-aware system

In phrase-based translation, source phrases are translated irrespective of their (source) context. This is often not perceived as a limitation as (i) typical text domains usually contain only few senses for polysemous words, thus limiting the use of word sense disambiguation (WSD); and (ii) using long-span target language models (4-grams and more) often capture sufficient context to select the more appropriate translation for a source phrase based on the target context. In fact, attempts at using source contexts in phrase-based SMT have to date failed to show important gains on standard evaluation test sets (Carpuat and Wu, 2007; Stroppa et al., 2007; Gimpel and Smith, 2008; Max et al., 2008). Importantly, in all conditions where gains have been obtained, the target language was the "morphologically-poor" English.

Nonetheless, there seems to be a clear consensus on the importance of better exploiting source contexts in SMT, so as to improve *phrase disambiguation*. The following sentence extract from the devtest corpus is a typical example where the lack of context in our phrase-based system yields an incorrect translation:
Source: *the long weekend comes with a price ...*
Target: *Le long week-end vient avec un prix ...*
(*the long weekend comes accompanied by a price*)

While grammatically correct, the French translation sounds unnatural, and getting the correct meaning requires knowledge of the idiom in the source language. In such a situation, the right context of the phrase *comes with* can be successfully used to propose a better translation.[3]

From an engineering perspective, integrating context into phrase-based SMT systems can be performed by (i) transforming source words into unique tokens, so as to record the original context of each entry of the phrase table; and by (ii) adding one or several contextual scores to the phrase table. Using standard MERT, the corresponding weights can be optimized on development data.

A typical contextual score corresponds to $p(\mathbf{e}|\mathbf{f}, C(\mathbf{f}))$, where $C(\mathbf{f})$ is some contextual information about the source phrase $\mathbf{f}$. An external disambiguation system can be used to provide one global context score (Stroppa et al., 2007; Carpuat and Wu, 2007; Max et al., 2008)); alternatively, several scores based on single features can be estimated using relative frequencies (Gimpel and Smith, 2008):

$$p(\mathbf{e}|\mathbf{f}, C(\mathbf{f})) = \frac{count(\mathbf{e}, \mathbf{f}, C(\mathbf{f}))}{\sum_{\mathbf{e}'} count(\mathbf{e}', \mathbf{f}, C(\mathbf{f}))}$$

For these experiments, we followed the latter approach, restricting ourselves to features representing the local context up to a fixed distance d (using the values 1 and 2 in our experiments) from the source phrase $\mathbf{f}_{start}^{end}$:

- lexical context features:
  - left context: $p(\mathbf{e}|\mathbf{f}, \mathbf{f}_{start-d}^{start-1})$
  - right context: $p(\mathbf{e}|\mathbf{f}, \mathbf{f}_{end+1}^{end+d})$
- shallow syntactic features (denoting t_1^F the sequence of POS tags for the source sentence):
  - left context: $p(\mathbf{e}|\mathbf{f}, t_{start-d}^{start-1})$
  - right context: $p(\mathbf{e}|\mathbf{f}, t_{end+1}^{end+d})$

As in (Gimpel and Smith, 2008), we filtered out all translations for which $p(\mathbf{e}|\mathbf{f}) < 0.0002$. This was necessary to make score computation practical given our available hardware resources.

Results on the devtest corpus for English→French were similar for the context-aware phrase-based and the baseline phrase-based system; small gains were achieved in the reverse direction (see Table 2). The same trend was observed on the test data.

Manual inspection of the output of the baseline and context-aware systems on the devtest corpus for English→French translation confirmed two facts: (1) performing phrase translation disambiguation is only useful if a more appropriate translation has been seen during training ; and (2) phrase translation disambiguation can capture important source dependencies that the target language model can not recover. The following ex-

[3] Our context-aware phrase-based system indeed proposes the appropriate translation: *Le long week-end a un prix.*

ample, involving an unseen sense[4] (*ball* in the semantic field of *dance* rather than *sports*), illustrates our first remark:
Source: *about 500 people attended the* ***ball*** .
Baseline : *Environ 500 personnes ont assisté à la* ***balle***.
+Context: *Environ 500 personnes ont participé à la* ***balle***.

The next example is a case where contextual information helped selecting an appropriate translation, in constrast to the baseline system.
Source: . . . *the new method for calculating pensions* ***due to begin*** *next year* . . .
Baseline : . . . *le nouveau mode de calcul des pensions* ***due à commencer*** *l'année prochaine* . . .
+Context: . . . *la nouvelle méthode de calcul des pensions* ***qui va débuter*** *l'année prochaine* . . .

3.2 Preliminary experiments with the GigaWord parallel corpus

One exciting novelty of this year's campaign was the availability of a very large parallel corpus for the en:fr pair, containing about 20M aligned sentences.

Our preliminary work consisted in selecting the most useful pairs of sentences, based on their average perplexity, as computed on our development language models. The top ranking sentences (about 8M sentences) were then fed into the usual system development procedure: alignment, reordering (for the N-code system), phrase pair extraction, model estimation. Given the unusual size of this corpus, each of these steps proved extremely resource intensive, and, for some systems, actually failed to complete. Contrarily, the N-code systems, conceptually simpler, proved to scale nicely.

Given the very late availability of this corpus, our experiments were very limited and we eventually failed to deliver the test submissions of our "GigaWord" system. Preliminary experiments using the N-code systems (see Table 2), however, showed a clear improvement of performance. There is no reason to doubt that similar gains would be observed with the Moses systems.

3.3 Experiments

The various systems presented above were all developed according to the same procedure: training used all the available parallel text; tuning was performed on dev2009a (1000 sentences), and our internal tests were performed on dev2009b (1000 sentences). Results are reported in table 2.

[4]This was confirmed after careful inspection of the phrase tables of the baseline system.

	en → fr		fr → en	
	Moses	Ncode	Moses	Ncode
small LM	20.06	18.98	21.14	20.41
Large LM	22.93	21.95	22.20	22.28
+context	23.06		22.69	
+giga		23.21		23.14

Table 2: Results on the devtest set

Our primary submission corresponds to the +context entry, our first contrast to Moses+LargeLM, and our second contrast to Ncode+largeLM. Due to lack of time, no official submission was submitted for the +giga variant. For the record, the score we eventually obtained on the test corpus was 26.81, slightly better than our primary submission which obtained a score of 25.74 (all these numbers were computed on the complete test set).

4 Conclusion

In this paper, we presented our statistical MT systems developed for the WMT'09 shared task. We used last year experiments to build competitive systems, which greatly benefited from in-house normalisation and language modeling tools.

One motivation for taking part in this campaign was to use the GigaWord corpus. Even if time did not allow us to submit a system based on this data, it was a interesting opportunity to confront ourselves with the technical challenge of scaling up our system development tools to very large parallel corpora. Our preliminary results indicate that this new resource can actually help improve our systems.

Naturally, future work includes adapting our systems so that they can use models learnt from corpora of the size of the GigaWord corpus. In parallel, we intend to keep on working on context-aware systems to study the impact of more types of scores, e.g. based on grammatical dependencies as in (Max et al., 2008). Given the difficulties we had tuning our systems, we feel that a preliminary task should be improving our tuning tools before addressing these developments.

Acknowledgments

This work was partly realised as part of the Quaero Program, funded by OSEO, the French agency for innovation.

References

M. Carpuat and D. Wu. 2007. Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In *Proceedings of Machine Translation Summit XI*, pages 73–80, Copenhagen, Denmark.

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

S. F. Chen and J. T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, NM.

J. M. Crego and J. B. Mariño. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

D. Déchelotte, G. Adda, A. Allauzen, O. Galibert, J.-L. Gauvain, H. Meynard, and F. Yvon. 2008. Limsi's statistical translation systems for WMT'08. In *Proceedings of the NAACL-HTL Statistical Machine Translation Workshop*, pages 107-100, Columbus, Ohio.

K. Gimpel and N. A. Smith. 2008. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, Columbus, Ohio.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*, Prague, Czech Republic.

A. Max, R. Makhloufi, and P. Langlais. 2008. Explorations in using grammatical dependencies for contextual phrase translation disambiguation. In *Proceedings of EAMT, poster session*, Hamburg, Germany.

J. B. Mariño, R. E. Banchs R, 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.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

N. Stroppa, A. van den Bosch, and A. Way. 2007. Exploiting source similarity for SMT using context-informed features. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*, pages 231–240, Skövde, Sweden.