LIUM's SMT Machine Translation Systems for WMT 2011

Holger Schwenk, Patrik Lambert, Loïc Barrault,
Christophe Servan, Haithem Afli and Sadaf Abdul-Rauf
LIUM, University of Le Mans
72085 Le Mans cedex 9, FRANCE
FirstName.LastName@lium.univ-lemans.fr

Abstract

This paper describes the development of French–English and English–French statistical machine translation systems for the 2011 WMT shared task evaluation. Our main systems were standard phrase-based statistical systems based on the Moses decoder, trained on the provided data only, but we also performed initial experiments with hierarchical systems. Additional, new features this year include improved translation model adaptation using monolingual data, a continuous space language model and the treatment of unknown words.

1 Introduction

This paper describes the statistical machine translation systems developed by the Computer Science laboratory at the University of Le Mans (LIUM) for the 2011 WMT shared task evaluation. We only considered the translation between French and English (in both directions). The main differences with respect to previous year's system (**?**) are as follows: use of more training data as provided by the organizers, improved translation model adaptation by unsupervised training, a continuous space language model for the translation into French, some attempts to automatically induce translations of unknown words and first experiments with hierarchical systems. These different points are described in the rest of the paper, together with a summary of the experimental results showing the impact of each component.

2 Resources Used

The following sections describe how the resources provided or allowed in the shared task were used to train the translation and language models of the system.

2.1 Bilingual data

Our system was developed in two stages. First, a baseline system was built to generate automatic translations of some of the monolingual data available. These automatic translations were then used directly with the source texts to build additional bitexts. In a second stage, these additional bilingual data were incorporated into the system (see Section 5 and Tables 4 and 5).

The latest version of the News-Commentary (NC) corpus and of the Europarl (Eparl) corpus (version 6) were used. We also took as training data a subset of the French–English Gigaword (10^9) corpus. We applied the same filters as last year to select this subset. The first one is a lexical filter based on the IBM model 1 cost (**?**) of each side of a sentence pair given the other side, normalized with respect to both sentence lengths. This filter was trained on a corpus composed of Eparl, NC, and UN data. The other filter is an n-gram language model (LM) cost of the target sentence (see Section 3), normalized with respect to its length. This filter was trained with all monolingual resources available except the 10^9 data. We generated two subsets, both by selecting sentence pairs with a lexical cost inferior to 4, and an LM cost respectively inferior to 2.3 (10^9_1, 115 million English words) and 2.6 (10^9_2, 232 million English words).

2.2 Use of Automatic Translations

Available human translated bitexts such as the Europarl or 10^9 corpus seem to be out-of domain for this task. We used two types of automatically extracted resources to adapt our system to the task domain.

First, we generated automatic translations of the provided monolingual News corpus and selected the sentences with a normalized translation cost (returned by the decoder) inferior to a threshold. The resulting bitext contain no new translations, since all words of the translation output come from the translation model, but it contains new combinations (phrases) of known words, and reinforces the probability of some phrase pairs (**?**). This year, we improved this method in the following way. In the original approach, the automatic translations are added to the human translated bitexts and a complete new system is build, including time consuming word alignment with GIZA++. For WMT'11, we directly used the word-to-word alignments produced by the decoder at the output instead of GIZA's alignments. This speeds-up the procedure and yields the same results in our experiments.

Second, as in last year's evaluation, we automatically extracted and aligned parallel sentences from comparable in-domain corpora. We used the AFP and APW news texts since they are available in the French and English LDC Gigaword corpora. The general architecture of our parallel sentence extraction system is described in details by Abdul-Rauf and Schwenk (**?**). 91M words from French into English using our first stage SMT system. These English sentences were then used to search for translations in the English AFP and APW texts of the Gigaword corpus using information retrieval techniques. The Lemur toolkit (**?**) was used for this purpose. The search was limited to a window of ±5 days from the date of the French news text. The retrieved candidate sentences were then filtered using the Translation Error Rate (TER) with respect to the automatic translations. In this study, sentences with a TER below 75% were kept. Sentences with a large length difference (French versus English) or containing a large fraction of numbers were also discarded. By these means, about 27M words of additional bitexts were obtained.

2.3 Monolingual data

The French and English target language models were trained on all provided monolingual data. In addition, LDC's Gigaword collection was used for both languages. Data corresponding to the development and test periods were removed from the Gigaword collections.

2.4 Development data

All development was done on *newstest2009*, and *newstest2010* was used as internal test set. The default Moses tokenization was used. However, we added abbreviations for the French tokenizer. All our models are case sensitive and include punctuation. The BLEU scores reported in this paper were calculated with the multi-bleu.perl tool and are case sensitive.

3 Architecture of the SMT system

The goal of statistical machine translation (SMT) is to produce a target sentence e from a source sentence f. Out main system is a phrase-based system (**?**; **?**), but we have also performed some experiments with a hierarchical system (**?**). Both use a log linear framework in order to introduce several models explaining the translation process:

$$\begin{aligned} e^* &= \arg\max p(e|f) \\ &= \arg\max_e \{exp(\sum_i \lambda_i h_i(e,f))\} \end{aligned} \quad (1)$$

The feature functions h_i are the system models and the λ_i weights are typically optimized to maximize a scoring function on a development set (**?**). The phrase-based system uses fourteen features functions, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty and a target language model (LM). The hierarchical system uses only 8 features: a LM weight, a word penalty and six weights for the translation model.

Both systems are based on the Moses SMT toolkit (**?**) and constructed as follows. First, word alignments in both directions are calculated. We used a multi-threaded version of the GIZA++ tool (**?**).[1]. This speeds up the process and corrects an error of GIZA++ that can appear with rare words.

Phrases, lexical reorderings or hierarchical rules are extracted using the default settings of the Moses toolkit. The parameters of Moses were tuned on *newstest2009*, using the 'new' MERT tool. We repeated the training process three times, each with a different seed value for the optimization algorithm. In this way we have an rough idea of the error introduced by the tuning process.

4-gram back-off LMs were used. The word list contains all the words of the bitext used to train the translation model and all words that appear at least ten times in the monolingual corpora. Words of the monolingual corpora containing special characters or sequences of uppercase characters were not included in the word list. Separate LMs were built on each data source with the SRI LM toolkit (**?**) and then linearly interpolated, optimizing the coefficients with an EM procedure. The perplexities of these LMs were 99.4 for French and 129.7 for English. In addition, we built a 5-gram continuous space language model for French (**?**). This model was trained on all the available French texts using a resampling technique. The continuous space language model is interpolated with the 4-gram back-off model and used to rescore n-best lists. This reduces the perplexity by about 8% relative.

4 Treatment of unknown words

Finally, we propose a method to actually add new translations to the system inspired from (**?**). To do so, we propose to identity unknown words and find possible translations for them.

Moses has two options when encountering an unknown word in the source language: keep it as it is or drop it. The first option may be a good choice for languages that use the same writing system since the unknown word may be a proper name. The second option is generally used when translating between language based on different scripts, *e.g.* translating from Arabic to English. Alternatively, we propose to automatically infer possible translations when translating from a morphologically rich language to a simpler language. In our case we use this approach to translate from French to English.

Several of the unknown words are actually adjectives, nouns or verbs in a particular form that is not known by the system, but the phrase table does contain the translation of a different form. As an example, the system may know how to translate the French word *finis* (masculine plural form) whereas it is unable to translate the word *finies* (female plural form) because it has not been encountered in the bitexts, while the translation of both word produce the english word *finished* .

After stemming, we may be able to find the translation in a dictionary which is automatically extracted from the phrase-table (see Table 1). This idea was already outlined by (**?**) to translate from Czech to English.

Source language French	Source language stemmed form	Target language English
finies	fini	finished
effacés	effacé	erased
hawaienne	hawaien	Hawaiian
...	...	...

Table 1: Example of translations from French to English which are automatically extracted from the phrase-table with the stemmed form.

First, we automatically extract a dictionary from the phrase table. This is done, be detecting all 1-to-1 entries in the phrase table. When there are multiple entries, all are kept with their lexical translations probabilities. Our dictionary has about 680k unique source words with a total of almost 1M translations.

The detection of unknown words is performed by comparing the source and the target segment in order to detect identical words. Once the unknown word is selected, we are looking for its stemmed form in the dictionary and propose some translations for the unknown word based on lexical score of the phrase table (see Table 2 for some examples). The stemmer used is the snowball stemmer[2]. Then the different hypothesis are evaluated with the target language model.

**** HS: I propose to delete this since I believe that it is wrong: you also need to consider all the th n-**

[1] The source is available at `http://www.cs.cmu.edu/~qing/`

[2] `http://snowball.tartarus.org`

source segment	les travaux sont **finis**
target segment	works are **finis**
stemmed word found	**fini**
translations found	**finished, ended**
segment proposed	works are **finished** works are **ended**
segment kept	works are **finished**

Table 2: Example of the treatment of an unknown French word and its automatically inferred translation.

corpus	newstest2010	subtest2010
number of sentences	2489	109
number of words	70522	3586
number of UNK detected	118	118
nbr of sentences containing UNK	109	109
BLEU Score without UNK process	29.43	24.31
BLEU Score with UNK process	29.43	24.33
TER Score without UNK process	53.08	58.54
TER Score with UNK process	53.08	58.59

Table 3: Statistics of the unknown word (UNK) processing algorithm on our internal test (newstest2010) and its sub-part containing only the processed sentences (subtest2010).

grams with the modified word in the CONTEXT. ** CS: I agree, by the way when I considered the whole sentence it gives the same results on each process. Here is a new suggestion :
Given this n-gram LM, the hypothesis re-evaluation considers the context of the whole sentence.

We processed the produced translations with this method. It can happen that some words are translations of themselves, e.g. the French word ”duel” can be translated by the English word ”duel”. If theses words are present into the extracted dictionary, we keep them. If we do not find any translation in our dictionary, we keep the translation. By these means we hope to keep named entities.

Several statistics made on our internal test (newstest2010) are shown in Table 3. Its shows that the influence of the detected unknown words is minimal. Only 0.16% of the words in the corpus are actually unknown. However, the main goal of this process is to increase the human readability and usefullness without degrading automatic metrics. We also expect a larger impact in other tasks for which we have smaller amounts of parallel training data. In future versions of this detection process, we will try to detect unknown words before the translation process and propose alternatives hypothesis to the Moses decoder.

5 Results and Discussion

The results of our SMT system for the French–English and English–French tasks are summarized in Tables 4 and 5, respectively. The MT metric scores are the average of three optimisations performed with different seeds (see Section 3). The numbers in parentheses are the standard deviation of these three values. The standard deviation gives a lower bound of the significance of the difference between two systems. If the difference between two average scores is less than the sum of the standard deviations, we can say that this difference is not significant. The reverse is not true. Note that most of the improvements shown in the tables are small and not significant. However many of the gains are cumulative and the sum of several small gains makes a significant difference.

Baseline French–English System

The first section of Table 4 shows results of the development of the baseline SMT system, used to generate automatic translations.

Although no French translations were generated, we did similar experiments in the English–French direction (first section of Table 5).

In both cases the best system is the one trained on the Europarl, News-comentaries and 10^9_2 corpora. This system was used to generate the automatic translations. We did not observe any gain when adding the United Nations data, so we discarded these data.

Impact of the Additional Bitexts

With the baseline French–English SMT system (see above), we translated the French News corpus to generate an additional bitext (News). We also translated some parts of the French LDC Gigaword corpus, to serve as queries to our IR system (see section 2.2). The resulting additional bitext is referred to as IR. The second section of Tables 4 and 5 summarize the system development including the additional bitexts.

With the News additional bitext added to Eparl+NC, we obtain a system of similar performance as the baseline system used to generate the automatic translations, but with less than half of the data. Adding the News corpus to a larger corpus, such as Eparl+NC+10^9_2, has less impact but still yields some improvement: 0.1 BLEU point in French–English and 0.3 in English–French. Thus, the News bitext translated from French to English may have more impact when translating from English to French than in the opposite direction. With the IR additional bitext added to Eparl+NC+10^9_2, we observe no improvement in French to English, and a very small improvement in English to French. However, added to the baseline system (Eparl+NC+10^9_2) adapted with the News data, the IR additional bitexts yield a small (0.2 BLEU) improvement in both translation directions.

Final System

In both translation directions our best system was the one trained on Eparl+NC+10^9_2+News+IR. We further achieved small improvements by pruning the phrase-table and by increasing the beam size. To prune the phrase-table, we used the 'sigtest-filter' available in moses (**?**), more precisely the $\alpha - \epsilon$ filter[3].

6 Conclusions and Further Work

We presented the development of our machine translation system for the French–English and English–French 2011 WMT shared task. Lessons learned this year include ....

Acknowledgments

This work has been partially funded by the European Union under the EuroMatrix Plus project IST-2007.2.2-FP7-231720 and the French project COSMAT ANR-09-CORD-004.

[3] The p-value of two-by-two contingency tables (describing the degree of association between a source and a target phrase) is calculated with Fisher exact test. This probability is interpreted as the probability of observing by chance an association that is at least as strong as the given one, and hence as its significance. An important special case of a table occurs when a phrase pair occurs exactly once in the corpus, and each of the component phrases occurs exactly once in its side of the parallel corpus (1-1-1 phrase pairs). In this case the negative log of the p-value is $\alpha = logN$ (N is number of sentence pairs in the corpus). $\alpha - \epsilon$ is the largest threshold that results in all of the 1-1-1 phrase pairs being included.

Bitext	#Fr Words	PT size	newstest2009	newstest2010		
	(M)	(M)	BLEU	BLEU	TER	METEOR
Eparl+NC	56	7.1	26.74	27.36 (0.19)	55.11 (0.14)	60.13 (0.05)
Eparl+NC+10^9_1	186	16.3	27.96	28.20 (0.04)	54.46 (0.10)	60.88 (0.05)
Eparl+NC+10^9_2	323	25.4	28.20	28.57 (0.10)	54.12 (0.13)	61.20 (0.05)
Eparl+NC+news	140	8.4	27.31	28.41 (0.13)	54.15 (0.14)	61.13 (0.04)
Eparl+NC+10^9_2+news	406	25.5	27.93	28.70 (0.24)	54.12 (0.16)	61.30 (0.20)
Eparl+NC+10^9_2+IR	351	25.3	28.07	28.51 (0.18)	54.07 (0.06)	61.18 (0.07)
Eparl+NC+10^9_2+news+IR	435	26.1	27.99	28.93 (0.02)	53.84 (0.07)	61.46 (0.07)
+larger beam+pruned PT	435	8.2	28.44	29.05 (0.14)	53.74 (0.16)	61.68 (0.09)

Table 4: French–English results: number of French words (in million), number of entries in the filtered phrase-table (in million) and BLEU scores in the development (newstest2009) and internal test (newstest2010) sets for the different systems developed. The BLEU scores and the number in parentheses are the average and standard deviation over 3 values (see Section 3.)

Bitext	#En Words	newstest2009	newstest2010	
	(M)	BLEU	BLEU	TER
Eparl+NC	52	26.20	28.06 (0.22)	56.85 (0.08)
Eparl+NC+10^9_1	167	26.84	29.08 (0.12)	55.83 (0.14)
Eparl+NC+10^9_2	284	26.95	29.29 (0.03)	55.77 (0.19)
Eparl+NC+10^9_2+news	299	27.34	29.56 (0.14)	55.44 (0.18)
Eparl+NC+10^9_2+IR	311	27.14	29.43 (0.12)	55.48 (0.06)
Eparl+NC+10^9_2+news+IR	371	27.42	29.73 (0.21)	55.16 (0.20)
+larger beam+pruned PT	371	27.49	29.82 (0.12)	55.32 (0.05)
+rescoring with CSLM	371		30.04	54.79

Table 5: English–French results: number of English words (in million) and BLEU scores in the development (newstest2009) and internal test (newstest2010) sets for the different systems developed. The BLEU scores and the number in parentheses are the average and standard deviation over 3 values (see Section 3.)

Bitext	#En Words	Rule-Table size	newstest2009		newstest2010	
	(M)	(M)	BLEU	TER	BLEU	TER
Eparl+NC	52		26.12	59.34	27.92	56.99
Eparl+NC+10^9_1	167		26.67	58.93	29.28	55.89
Eparl+NC+10^9_2	284		26.85	58.53	29.12	55.71

Table 6: English–French results: number of English words (in million) and BLEU scores in the development (newstest2009) and internal test (newstest2010) sets for the different systems developed.

Bitext	#Fr Words	Rule-Table size	newstest2009		newstest2010	
	(M)	(M)	BLEU	TER	BLEU	TER
Eparl+NC	56		26.87	56.38	27.88	54.73
Eparl+NC+10^9_1	186		27.69	55.60	28.37	54.33

Table 7: French–English results: number of French words (in million) and BLEU scores in the development (newstest2009) and internal test (newstest2010) sets for the different systems developed.