LIUM's Statistical Machine Translation System for IWSLT 2010

Anthony Rousseau, Loïc Barrault, Paul Deléglise, Yannick Estève

Laboratoire Informatique de l'Université du Maine (LIUM)
University of Le Mans, France

firstname.lastname@lium.univ-lemans.fr

Abstract

This paper describes the two systems developed by the LIUM laboratory for the 2010 IWSLT evaluation campaign. We participated to the new English to French TALK task. We developed two systems, one for each evaluation condition, both being statistical phrase-based systems using the the Moses toolkit. Several approaches were investigated.

1. Introduction

This paper describes the systems developed by the LIUM laboratory for the 2010 IWSLT evaluation campaign. This year, a new task (named TALK task) has been proposed, which is based on the TED talks and consists in translating the talks transcriptions from English to French. For this evaluation, two submissions were required, as two input conditions for translation were proposed. The first, called Correct Recognition Results (CRR) condition, uses the Automatic Speech Recognition (ASR) manual transcription as input, and the second, called ASR condition, uses an automatic speech recognition output. Thus we developed two specific systems, one for each input condition.

The remainder of this paper is structured as follows: in section 2, we describe the individual systems setup and the specific strategies for translating in ASR condition. Particular approaches and issues like the use of ASR lattices or segmentation problems are discussed in section 3, then experimental results are summarized in section 4. The paper concludes with a discussion on future research issues in section 5.

2. SMT Systems

Since the proposed TALK task was divided in two parts for submission (CRR and ASR conditions), we developed two different systems with some specificities regarding the ASR one. This section focuses on data preparation for the different systems, then on language modeling and eventually on the description of our two systems.

2.1. Available resources

The organizers of IWSLT provide several specific corpora that can be used to train and optimize the translation systems. The characteristics of these corpora are summarized in Table 1. The translation models were trained on selected bitext corpora among the proposed ones. The target language model was trained on the French side of the those corpora. No additional texts were used (*constrained condition*).

corpus	#lines	#tok English	#tok French
TED v1.1	84.5k	877k	943k
News-Commentary 10	84.6k	2M	2.4M
Europarl v5	1.6M	45M	45M
UN200x	7.2M	211.7M	240.2M
Gigaword release 2	22.5M	662.7M	771.7M
TED dev CRR	1307	12554	12528
TED dev ASR 1Best	259	11334	n/a
TED test CRR	3502	31980	n/a
TED test ASR 1Best	758	28115	n/a

Table 1: Characteristics of the provided bitext corpora.

2.2. Data preprocessing

The data proposed for this task consists of a huge amount of text, with a total number of tokens close to one billion. Moreover, some of these corpora, like the Gigaword corpus, are very noisy and contain many irrelevant data.

In order to improve the performance of the systems, we considered some processing aimed at increasing the quality of the data.

Thus, after a classical tokenization using the tokenizer provided in the Moses toolkit, we filtered the lines of the biggest bitexts (Europarl, UN200x and Gigaword) with a method using a lexical model. This lexical model is trained on the same corpora than the proposed ones for the task. We first calculated the lexical cost of the translation for each segment, then we applied a threshold on these lexical costs in order to extract the lines with the lower cost, as this method was proven successful in a recent work from LIUM [1].

As we can see in Table 2, and as we expected, we greatly reduced the size of the noisiest corpus, which is Gigaword, thus we can assume that most of the irrelevant data has been removed.

corpus	#unfiltered lines	#filtered lines
Europarl.v5	1.6M	1.5M
UN200x	7.2M	7.1M
Gigaword_fr-en	22.5M	12.8M

Table 2: Comparison between filtered and unfiltered data.

2.3. Language Modeling

The language models (LM) used for the task were trained using the SRILM toolkit [2]. The final LM is a 4-gram back-off (Kneser-Ney discounting) target language model built on all available French data.

In order to select the optimal vocabulary, we trained unigram models and we interpolated them to get a global unigram model. That model is then sorted according to the word probabilities, which allows us to select the more probable words appearing in the corpora [3]. Starting from this vocabulary, we constructed a 4-gram LM for each corpus, which are then interpolated to obtain the final LM. The interpolation weights are optimized on the development corpus with a numerical optimizer. As we expected, the TED corpus is the one with the biggest weight in the final LM. Different sizes of vocabulary (150k, 450k, all) were used to generate several final LMs in order to measure the impact of vocabulary size on translation quality.

2.4. SMT system for CRR condition

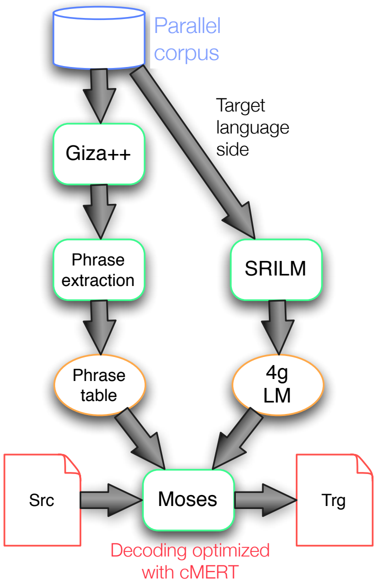

Figure 1: Architecture of our SMT system for CRR condition.

Our statistical phrase-based systems are based on the Moses SMT toolkit [4] and constructed as follows. First, Giza++ is used to perform word alignments in both directions. Second, phrases and lexical reorderings are extracted. Both steps use the default settings of the Moses SMT toolkit. In our systems, fourteen features functions were used, 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. The coefficients of these feature functions are tuned on the development corpus with the cMERT tool using 100-best lists. The Figure 1 shows the basic architecture of our SMT system.

2.4.1. Data selection

As we said in section 2.2, the amount of data in the proposed corpora is very huge, thus it may be beneficial to select a subset of better quality, by limiting the training data to some of the available corpora. We empirically determined which corpora were the best for this task by generating several systems, each using only a subset of corpora, optimizing them on the development corpus and comparing the resulting BLEU scores.

The Table 3 shows the BLEU scores obtained on different corpus combinations. Regarding these experiments, the best combination we could determine was to select the TED corpus, along with the news-commentary and the Europarl ones, for a total amount of data of about forty-seven millions of tokens.

corpus combination	BLEU score	#lines	#tokens (French)
All corpora	26.14	21.7M	672.4M
TED + NC + Eparl	**26.57**	**1.7M**	**47.4M**
TED + NC + UN200x	25.98	7.3M	217.0M
All except Gigaword	25.87	8.8M	261.5M

Table 3: Comparison of BLEU scores obtained on development corpus for various corpus combinations.

2.5. SMT system for ASR condition

Translating Automatic Speech Recognition outputs with a SMT system requires some adaptation to the specificities of ASR hypotheses. Indeed, ASR outputs by default are lowercased with no punctuation, since classic ASR evaluation does not take into account these particularities. Besides that, text normalization between ASR and SMT differs on some points, for instance the numbers, which are written in letters for ASR, the contractions as well as acronyms. In order to get a somewhat satisfying score while translating text in ASR condition, these issues must be addressed.

Our approach for treating those particularities is the following. We first processed the parallel corpora to obtain training data which resembles to ASR condition text, meaning that we suppressed all punctuation, lowercased all words (except proper nouns and words which are always capitalized, like the word “I” in English), transformed numbers into letters, normalized many contractions (like “Mr.” into “mister” or “it is” into “it’s”) and symbols for ASR. Then we trained a system using these modified corpora, with the same corpus combination than we used for our CRR system. We then optimized the newly-trained system on the provided 1-best development corpus. Besides that, a specific language model was trained with no punctuation nor case, also using the target language side of the modified bitexts. This led to better experimental results on the development corpus (BLEU = 18.49) than directly translating ASR condition text with the CRR system (BLEU = 15.86).

2.5.1. Case and punctuation

Regarding the case and punctuation issues, we adopted a specific approach, which is only based on the target language side. We considered the original French corpus and the corresponding “ASR condition” corpus as bitexts and trained a new system with these parallel data.

During this training, it is preferable to limit the lexical reordering, in order to avoid unwanted modifications to our hypothesis. The idea behind this is to consider the French in ASR condition as a full-fledged language, consequently our recaser can be regarded as a “French ASR-to-French” SMT system. This system was then optimized on the CRR development corpus, which is nothing more than the ASR manual transcription, except that it contains case and punctuation. The Figure 2 presents the global architecture of our SMT system for ASR condition.

3. Translating ASR outputs

3.1. Handling ASR word lattices

For this evaluation campaign, word lattices, n-best lists and 1-best hypotheses were available. The use of word lattices as input of Moses system is described in [5]. Alternatively, we decided to generate another kind of input, namely confusion networks [6], computed from the word lattices.

The word lattices were provided by the organizers under SLF format, which is the file format used by the HTK tools. In practice, word lattices provided by the organizers were very large, too large to be reasonably managed “as is” by the Moses decoder. This large size can be mainly explained because the word lattice topology strictly represents the history constraints applied to words in these word lattices, making their language model scores consistent with their history. No information was provided by organizers about these word lattices, but it is possible to deduce that the word lattice topology was constrained by a 4-gram language model.

So, in order to use these word lattices as inputs of the Moses decoder, we had to reduce their size. To do this, we used the tools we have implemented in the LIUM ASR system [7] to manage word lattices, with some minor modifications to make these tools entirely adapted to this task. This can be summarize by the following steps:

1. First, link posteriors are computed using the *forward-backward* algorithm: this is relevant to handle link scores properly, for instance to merge some links associated to the same word.

2. Secondly, some words in the word lattices were split into several words in order to make the tokenization used in the word lattice closer to the one used in the translation and language models. To do this, we had to inject some new links in the graphs.

3. In step 3, the 4-gram topology of the graph was broken by merging links with identical words having different history but located in equivalent temporal area. When merging two links, their probability *a posteriori* are added.

4. Then a pruning process is made in order to remove links associated to low posteriors. The step 3 and this step are repeated two times.

5. Filler words, such as *hesitation* or *breath*, are deleted and ϵ (null transitions) are removed.

6. Last, word lattices are written into PLF format at one side, and are transformed into confusion networks in the other one.

Notice that some information were missing about the ASR word lattice building. For example, no information was given about the way in which the word insertion penalty used during the speech recognition process was integrated into the word lattice scores. Moreover, two linguistic weights were provided in the header of the SLF files, but certainly only one was applied. However, the word error rate of the best recognition hypotheses computed from the SLF word lattices on the development set (evaluated, of course, on the source language data) reaches 26.4% WER instead of the 24.8% WER obtained by the 1-best hypotheses distributed by the organizers. To get our best hypotheses, we used a word insertion penalty equals to 0.62 and a linguistic weight equals to 12.5.

3.2. Rescoring n-best SMT hypotheses with POS LM

It was recently shown that using morpho-syntactic post-processing on n-best ASR hypotheses should improve speech recognition for French language, especially by using high order part-of-speech (POS) n-gram language models [8]. We have applied this approach to rescore n-best SMT hypotheses with a 7-gram POS LM. To do this, we used the *lia_tagg* tool[1] to tag the n-best SMT hypotheses. This tool was also used to tag the French training data. We computed a 7-gram POS LM on the POS-tagged training data and applied it to

[1] *lia_tagg* was developed by Frédéric Béchet and is distributed by the LIA under GPL license

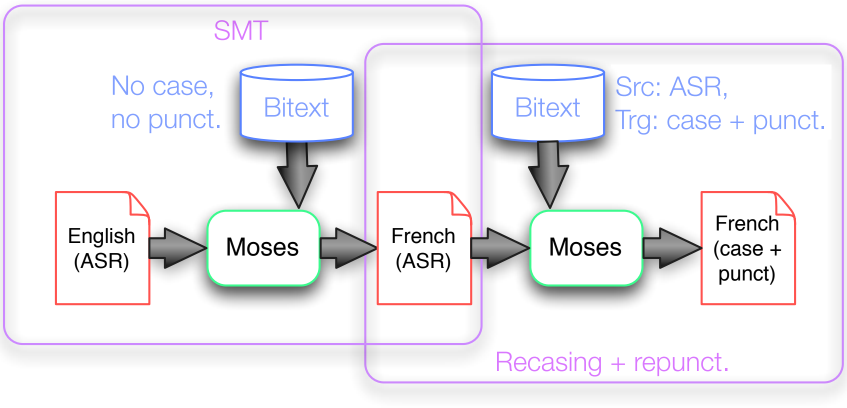

Figure 2: Global architecture of our SMT system for ASR condition.

give a POS LM score to each tagged n-best SMT hypothesis. This score was taken into consideration to recompute the global score of each n-best hypothesis with optimized linear coefficients.

	dev set	test set
Best point without POS	19.55	**20.98**
Best point after tuning	**19.79**	20.65

Table 4: BLEU scores obtained with n-best list rescoring using a 7-gram POS LM.

It seems that this approach does not generalize very well on test data. This is probably due to the fact that the 16 parameters are optimized with cMert on a small corpus (250 sentences), which lead to over-tuning.

However, this is a rather disappointing result which does not reflect what can be seen in literature. A deeper analysis of the tags obtained during this processing is necessary to clearly understand those results.

4. Experimental Results

4.1. Official results

The results of our systems for this evaluation campaign are presented in Table 5.

	dev set		test set	
	CRR condition - Eval condition 1			
	BLEU	TER	BLEU	TER
	26.45	61.02	25.07	57.60
1-Best	ASR condition - Eval condition 3			
	BLEU	TER	BLEU	TER
	18.49	70.01	18.27	70.92

Table 5: Official results for the TALK task in CRR and ASR conditions.

4.2. Two weeks later

Since the end of evaluation period, we got some better result by using the available word lattices with the approach detailed in section 3.1

As we can see in Table 6, while PLF and CN show higher WER, a better BLEU score can be obtain using them as input of the SMT system. This is also related to the fact that we did not have the value of the word insertion penalty nor those for the linguistic weights (see section 3.1).

Moreover, optimizing weights for CN is really tricky. Indeed, changing the value for the *weight-i*, the parameter used to weight the score on the edge in the CN, leads to changing the size of the input (as the epsilon transition will be more or less likely crossed), which is quite disturbing for the optimizer. Another point is that such a process is very slow because PLF and CN need a binary phrase table which does not take the most of multithreading (lots of mutual blocking).

Also the difference between official results and those proposed here lie in the fact that two different versions of Moses toolkit were used. The official results were obtained using Moses from February 2009 and the new ones come from Moses from August 2010.

5. Conclusion

For this year IWSLT evaluation, the LIUM participated in the new TALK task, consisting in translating ASR outputs from english to french.

Several results are worth mentioning. In one hand, rescoring n-best list with a part-of-speech language model for french provided some rather unexpected results. Tagger outputs, probabilities obtained with the POS LM and tuning process will be further investigated in the future.

In the other hand, the results obtained show that using bigger search space at the input of the SMT system leads to improvement regarding to using only the 1-best hypothesis from the ASR system. This, even when the WER obtained

	WER	ASR condition - Eval condition 1				ASR condition - Eval condition 3			
PLF	26.4	dev set		test set		dev set		test set	
		BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
		-	-	**18.48**	70.88	**19.44**	69.33	**20.98**	66.09
CN	26.1	dev set		test set		dev set		test set	
		BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
		-	-	-	-	19.39	69.39	-	-
1-Best	24.8	dev set		test set		dev set		test set	
		BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
		-	-	17.86	71.66	19.19	69.45	20.14	66.77

Table 6: Late results for TALK task in ASR conditions.

from the graph (using our own weights for word insertion penalty and language model) is higher than the WER provided by the organisers. However, in order to deal with the size and the amount of data, some processing are necessary in order to make the use of the graphs manageable as well as increasing the quality and the usefulness of the data.

6. References

[1] Lambert P., Abdul-Rauf S. and Schwenk H., "LIUM SMT Machine Translation System for WMT 2010", ACL Workshop on Statistical Machine Translation (WMT'10), Uppsala (Sweden), pp. 127-132, 2010.

[2] Stolcke A., "SRILM - An Extensible Language Modeling Toolkit", Proc. Intl. Conf. Spoken Language Processing, Denver (United-States), September 2002.

[3] Allauzen, A. and Gauvain, J.-L., "Construction automatique du vocabulaire d'un système de transcription", Journées d'Etude sur la Parole 2004, Fès (Maroc), 2004.

[4] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation", ACL demonstration session, 2007.

[5] C. Dyer, S. Muresan, and P. Resnik, "Generalizing Word Lattice Translation", Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2008.

[6] N. Bertoldi, R. Zens and M. Federico, "Speech Translation by Confusion Network Decoding", International Conference on Acoustics, Speech, and Signal Processing, pp. 1297-1300, 2007.

[7] Deléglise P., Estève Y., Meignier S., Merlin T., "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?", Interspeech 2009, Brighton (United Kingdom), 2009.

[8] Stéphane Huet, Guillaume Gravier, Pascale Sébillot. "Morpho-Syntactic Post-Processing with N-best Lists for Improved French Automatic Speech Recognition", Computer Speech and Language, 24(4):663-684, October 2010.