LIG English-French Spoken Language Translation System for IWSLT 2011

Benjamin LECOUTEUX, Laurent BESACIER, Hervé BLANCHON

LIG Laboratory
University of Grenoble, France
name.surname@imag.fr

Abstract

This paper describes the system developed by the LIG laboratory for the 2011 IWSLT evaluation. We participated to the English-French MT and SLT tasks.
The development of a reference translation system (MT task), as well as an ASR output translation system (SLT task) are presented. We focus this year on the SLT task and on the use of multiple 1-best ASR outputs to improve overall translation quality. The main experiment presented here compares the performance of a SLT system where multiple ASR 1-best are combined before translation (*source* combination), with a SLT system where multiple ASR 1-best are translated, the system combination being conducted afterwards on the target side (*target* combination). The experimental results show that the second approach (*target* combination) overpasses the first one, when the performance is measured with BLEU.

1. Introduction

This paper describes LIG approach for the evaluation campaign of the 2011 International Workshop on Spoken Language Translation (IWSLT-2011), English-French MT and SLT tasks.

This year we focus on the SLT task and on the use of multiple 1-best ASR outputs to improve translation. Two different approaches are proposed:

-source combination: multiple ASR 1-best are combined before translation,

-target combination: multiple ASR 1-best are translated, before applying system combination on the target side.

The remainder of the paper is structured as follows. Section 2 reminds the starting point of this work, namely the former LIG SLT system presented last year for IWSLT 2010. Then, we describe chronologically the work done this year to improve both MT and SLT English-French systems, including the update of the models with data provided this year (section 3). The best system obtained in section 3 is used for the experiments detailed in section 4 where *target* combination is compared to *source* combination. Finally, in section 5 we sum up our work.

2. Overview of MT and SLT LIG systems in 2010

This section describes the starting point of this work which is the LIG system presented last year for IWSLT 2010. More details on this system can be found in [1].

Last year, a new task was dedicated to the translation of the TED Talks corpus, a collection of public speeches on a variety of topics for which video, transcripts and translations are available on the Web. Training data for this exercise was limited to a supplied collection of freely available parallel texts, including a parallel corpus of TED Talks. The translation input conditions of the TALK task consisted of (1) automatic speech recognition (ASR) outputs, i.e., word lattices (SLF), N-best lists (NBEST) and 1-best (1BEST) speech recognition results, and (2) correct recognition results (CRR), i.e., text input without speech recognition errors. Participants of the TALK task had to submit MT runs for both input conditions.

2.1. Resources Used in 2010

Last year, we used the TED Talks collection plus other parallel corpora distributed by the ACL 2010 Workshop on Statistical Machine Translation (WMT).

For the training of the translation models, the provided Europarl and News parallel corpora were used (total 1,767,780 sentences) as well as the TED training corpus (total 47,652 sentences). For the language model training, in addition to the French side of the bitexts described above (News-mono+TED-mono), the 2010 News monolingual corpus in French was available (total 15,234,997 sentences).

The TED dev set (934 sentences) was used both for tuning and evaluation purpose. This corpus will be referred to as *Dev2010* in the rest of this paper.

2.2. Preprocessing / Post-processing in 2010

As preprocessing, we lowercased and tokenized all the data but kept punctuation for the LM and TM models training. Before translation, a source English sentence is thus lowercased and tokenized. The translated output in French needs to be detokenized and recased. The best technique found to re-case the translated output used a SMT-like approach where a phrase table was trained from a parallel French no-case/case corpus (trained on the News monolingual corpus in French of 15M sentences, see [1]).

For the Reference translation (MT) task, the punctuation of the translated output was refined using the punctuation of the source sentence (practically, the ending punctuation mark of the source sentence was put at the end of the translated sentence).

2.3. Language modeling in 2010

The target language model was a standard 3-gram language model trained using the SRI language modeling toolkit [7]. The smoothing technique applied was the modified Kneser-Ney discounting with interpolation.
We interpolated a LM trained on the TED training data (47k sentences) with a LM trained on Europarl, News, UN and

News-mono (24M sentences in total). After a perplexity test to optimize the interpolation weight (on Dev2010), we chose an interpolation weight equal to 0.5.

2.4. Translation modeling and tuning

For the translation model training, the uncased (but punctuated) corpus was word aligned and then, the pairs of source and corresponding target phrases were extracted from the word-aligned bilingual training corpus using the scripts provided with the Moses decoder [3]. The result is a phrase-table containing all the aligned phrases. This phrase-table, produced by the translation modeling, is used to extract several translations models. In the experiments reported here, only 8 features were used in the phrase-based models: 5 translation model scores, 1 distance-based reordering score, 1 LM score and 1 word penalty score.

We used the Minimum Error Rate Training (MERT) method to tune the weights. MERT was applied on the TED Dev2010 corpus (934 sentences). Moreover, it is important to note that, during tuning, punctuation was systematically removed from the Nbest lists and BLEU was calculated using un-punctuated references. While such tuning procedure might be sub-optimal to optimize BLEU (cased), we did this to anticipate the ASR output translation task for which decoding (and tuning) is also done without punctuation.

2.5. Other aspects of the LIG 2010 MT system

Last year, additional improvements over the above described baseline were proposed (see [1] for more details):

-do not reorder over punctuation during decoding,

-apply phrase-table pruning with a technique similar to [4] (retuning with MERT needed after pruning).

Table 1 reports the results obtained on Dev2010 (934 sentences) and Tst2010 (1664 sentences) with last year LIG system.

Table 1: Performance of the IWSLT 2010 LIG MT system using BLEU [5] – BLEU measured with punct+case (c+p), case only (c) and none (x)

Corpus	BLEU c+p	BLEU c	BLEU X
Dev2010	0.2408	0.2179	02311
Tst2010	0.2758	0.2479	0.2590

2.6. SLT system for IWSLT 2010

For the speech translation (SLT) task, the TM and LM models described above were used. However, the pre-/post-processing was different since, for instance, no "source punctuation" could be used in the case of ASR input.

First, in order to be consistent with our translation model, the ASR output was lowercased and tokenized before translation. Moreover, the (source) English ASR output was re-punctuated (see [1] for more details).

Finally, it was necessary to develop a true re-punctuation system for French in the case of ASR output translation. This was done by building a French language model trained on punctuated and uncased French data (Europarl +News+UN+Newsmono: 24M sentences in total). The punctuation was restored after translation using this LM and the hidden-ngram command from SRILM toolkit. After re-punctuation, we used the SMT-based recaser presented earlier. For the SLT task, the final system submitted by LIG in 2010 was ranked among the best sites that participated to the TALK task last year.

3. Improvements of MT and SLT systems done for 2011

3.1. Iterative improvement of the MT system

Table 2 summarizes the iterative improvements done this year over the LIG 2010 system. First, we evaluated the performance of a phrase-table trained on the TED 2011 bilingual data (107268 sentences in total) only with and without tuning (2,3). The target language model was also updated using the TED 2011 mono (111431 sentences) data (4), which slightly increased the performance. The results obtained show a reasonable performance of the PT trained on TED 2011 only, so we experimented multiple phrase-table decoding where translation options are collected from one table, and additional options are collected from the other table. When the same translation option (in terms of identical input phrase and output phrase) is found in multiple tables, separate translation options are created for each occurrence, but with different scores (this corresponds to the *either* option defined in the *moses* advanced features[1]) After retuning on dev2010 data, this approach improved the system by more than 1 point BLEU (5,6). Note that in this case there are 10 phrase table translation features instead of 5.

Table 2: Iterative improvement of the LIG MT system in 2011

System	BLEU c+p dev2010/ tst2010	BLEU c dev2010/ tst2010	BLEU x dev2010/ tst2010
1. LIG 2010 2010 bitexts	0.2408/ 0.2758	0.2179/ 0.2479	02311/ 0.2590
2. PT trained on TED2011 bitext only (no tuning)	0.2270/ 0.2782	0.2044/ 0.2508	0.2167/ 0.2611
3. PT trained on TED2011 bitext only (+tuning)	0.2411/ 0.2781	0.2168/ 0.2513	0.2296/ 0.2621
4. (1)+update LM using TED 2011 mono	0.2452/ 0.2789	0.2207/ 0.2516	0.2335/ 0.2623
5. Multiple PT - Either(1,4) - no tuning + updated LM	0.2397/ **0.2898**	0.2167/ 0.2618	0.2293/ 0.2726
6. Multiple PT - Either(1,4) + tuning + updated LM	**0.2524**/ 0.2896	**0.2289**/ **0.2623**	**0.2420**/ **0.2733**

3.2. Improvement of the SLT system

The pre-/-post- processing for SLT described in section 2.6 was not changed this year for 2011 evaluation. However, we performed a tuning adapted to ASR input by re-estimating

[1] http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc15

the log-linear weights using the dev2010 ASR output (corresponding to a *rover* between several systems, provided by the organizers). The BLEU score was improved significantly using the new weights both on dev2010 and tst2010. The other improvements of the SLT system are described in section 4 which details the source/target combination approaches.

Table 3: Iterative improvement of the LIG SLT system in 2011 (using the rover provided by the organizers as input)

Corpus	BLEU c+p dev2010/ tst2010	BLEU c dev2010/ tst2010	BLEU x dev2010/ tst2010
7. (6) + pre-/post-process described in 2.6	0.1670/ 0.2027	0.1606/ 0.1992	0.1709/ 0.2081
8. (7)+ tuning on ASR input (Dev2010)	**0.1745/ 0.2087**	**0.1671/ 0.2046**	**0.1766/ 0.2133**

4. Source versus Target Combination

This year, since several ASR system outputs were provided for the evaluation (see table 4 for an overview of ASR system performance on tst2010 data), we decided to investigate different combination techniques. More precisely, we compared the performance of a SLT system where multiple ASR 1-best are combined before translation (*source* combination), with a SLT system where multiple ASR 1-best are translated, the system combination being conducted afterwards on the target side (*target* combination). The TM and LM used, as well as the log-linear weights are the one of the system (8) described in section 3.2 (performance given in table 3). This means that the log-linear weights of the SMT system were not re-tuned in the experiments described in this section.

Table 4: ASR performance [2] of the system (outputs) used (on *tst2010*)

System	WER%
0	17.1
1	18.2
2	17.4
3 (not used)	27.3
4	**15.3**

4.1 Source combination

In order to combine sources we applied a classical ROVER [8] weighted by the ASR WER quality. The used cost function for word selection is:

*alpha*Sum(WordOcc) + (1-alpha)*Sum(Confidence(W))*

Where alpha=0.9 and confidence scores are empirically defined: 1 for best system (4), 0.8 for systems (2) and (0) and 0.5 for system (1).

4.2 Target combination

In that case, we propose a MT systems combination similar to the one used in [6]. System combination is based on the 500-best translated outputs generated from each ASR source system. We used the Moses option *distinct*, ensuring that the hypotheses produced for a given sentence are different inside an N-best list. Each N-best list is associated with a set of 13 features:

- 10 translation model scores (2 phrase tables * 5 scores each)
- 1 distance-based reordering score
- 1 language model score
- 1 word penalty score

N-best are combined in several steps. The first one takes as input lowercased 500-best lists, since preliminary experiments have shown a better behaviour using only lowercased output (with cased output, combination presents some degradations). The score combination weights are optimized on a development corpus, in order to maximize the BLEU score at the sentence level when N-best lists are reordered according to the 13 available scores. To this end, we resorted to the SRILM nbest-optimize tool to do a simplex-based Amoeba search [10] on the error function with multiple restarts to avoid local minima.

Once the optimized feature weights are computed independently for each ASR source, N-best lists are turned into confusion networks [9]. The 13 features are used to compute posteriors relatively to all the hypotheses in the N-best list.
Confusion networks are computed for each sentence and for each system. Then, these confusion networks computed for each sentence are merged into a single one. A ROVER is applied on the combined confusion network and generates a lowercased 1-best. The usual post-processing described in 2.6 is finally applied as usual to obtain adequate output.

On this system we observe a different behaviour compared to the one presented in [6]: combining the N-best of a single system does not improve the BLEU score. Thus, all the experiments reported below involves combination of several N-best lists (except for first four lines of table 5).

4.3 Experiments

The results obtained from individual ASR systems show that the best transcription system (system4) leads to the best BLEU score while the worst one (system1) leads to the lowest BLEU score (2.9% WER absolute difference gave 1.6 BLEU difference). However, the correlation between ASR performance and BLEU is not so clear while looking at results for system0 and system2 (lower WER for system 0 but lower BLEU too).

Table 5: Source vs Target Combination (system3 has been removed from the experiments) – **here combination tuned on tst2010 and evaluated on dev2010**

Combination	BLEU c+p dev2010/ tst2010	BLEU c dev2010/ tst2010	BLEU x dev2010/ tst2010
Sys 0 alone	0.1671/ 0.2012	0.1602/ 0.1957	0.1695/ 0.2039
Sys 1 alone	0.1608/ 0.1944	0.1534/ 0.1909	0.1622/ 0.1985
Sys 2 alone	0.1737/ 0.2027	0.1664/ 0.1975	0.1768/ 0.2072
Sys 4 alone	0.1770/ 0.2082	0.1709/ 0.2033	0.1811/ 0.2125
Target comb. (systems 42)	0.1772/ 0.2085	0.1710/ 0.2036	0.1812/ 0.2130
Source comb. (rover systems 420) done at LIG	0.1787/ 0.2139	0.1709/ 0.2099	0.1811/ 0.2191
Target comb. (systems 420)	0.1815/ 0.2136	0.1748/ 0.2087	0.1852/ 0.2178
Source comb. (rover systems 0213) provided by IWSLT orga. (cf tab 3)	*0.1745/ 0.2087*	*0.1671/ 0.2046*	*0.1766/ 0.2133*
Source comb. (rover systems 4021) done at LIG	0.1797/ 0.2159	0.1726/ 0.2115	0.1826/ 0.2209
Target comb. (systems 4021)	**0.1841**/ 0.2143	**0.1782**/ 0.2099	**0.1889**/ 0.2189
Source+Target comb. (systems 4021R)	0.1818/ **0.2166**	0.1758/ **0.2120**	0.1859/ **0.2215**

As far as system combination is concerned, it is important to note that we decided to tune the combination weights on tst2010 data, which is twice bigger than dev2010 data. Thus, dev2010 was considered as a validation test in the case of table 5 results.
When two systems are available, target combination is inefficient while source combination cannot be applied. When three systems are available, the target combination is clearly better than the source combination on the validation set (which is dev2010, *cf* remark above). The same trend is observed with four systems. We can note that as more ASR systems (2, 3, 4) are added to the combination, the overall performance improves.

So, in order to take advantage of both combinations we also experimented a source+target combination where the (source) rover is added as a new system to the target combination method. However in this last experiment a slight BLEU degradation is observed on the validation set (dev2010), even if the results on the development set (tst2010 here) are better. This disappointing result may be explained by the fact that the ROVER source introduces redundant information and leads to the elimination of marginal assumptions.

4.4 Official Results

At the time of submission, we had not evaluated all the combinations described in table 5. So, at this time, the source+target combined system (last line of table 5) was submitted as our "primary" (LIG_P) system, our contrastive system corresponding to source combination strategy only (LIG_C1 ; rover4021). The official results of table 6 (obtained on tst2011 data) confirm that the target combination (and source+target) outperforms the source combination.

Table 6 : Official automatic evaluation results obtained by LIG at IWSLT11 (BLEU score) – SLT Task

System.	bleu(p+c)	bleu(x)
LIG_P (Tst2011) source+target comb. (4201R)	0,2485	0,2598
LIG_C1 (Tst2011) source comb. (4201)	0,2453	0,2561
LIG_PostEval (Tst2011) Target comb (4201)	**0.2489**	**0.2599**

5. Conclusion

This paper presented the work done at LIG this year for IWSLT2011. While the English-French MT was mostly updated on the new data, without radical changes, we proposed several approaches to take advantage of multiple ASR system outputs. The experimental results obtained show that combining translation hypotheses (obtained from several translated ASR 1best) on the target language side lead to better results than combining ASR 1best on the source side, before translation (0.4 BLEU improvement observed).

6. References

[1] L. Besacier, H. Afli, T.N. Do, H. Blanchon, M. Potet "LIG Statistical Machine Translation Systems for IWSLT 2010" *IWSLT 2010*. Paris, France. December 2010.

[2] Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stueker - Overview of the IWSLT 2011 Evaluation Campaign - Proceedings of the International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA, 2011.

[3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". *ACL 2007*, demonstration session.

[4] H. Johnson & al (2007) "Improving Translation Quality by Discarding Most of the Phrasetable". In *proceedings of the EMNLP-CoNLL 2007*. pp 967-975.

[5] Papineni, K., Roukos, S., Ward, T., and Zhu, W., "BLEU: A method for automatic evaluation of machine translation", *ACL'02*, pp. 311-318, Philadelphia, USA, July 2002.

[6] M. Potet, R. Rubino, B. Lecouteux, S. Huet, L. Besacier, H. Blanchon, and F. Lefevre. The LIGA machine translation system for WMT 2011. In Proceedings EMNLP and ACL Workshop on Machine Translation (WMT), Edinburgh (Scotland), July 2011.

[7] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", *ICSLP'02*, vol. 2, pp. 901-904, Denver, Colorado, September 2002.

[8] Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates:recognizer output voting error reduction (ROVER). In Proceedings of the IEEE Workshop on Automatic Speech Recognition and

Understanding , pages 347–354, Santa Barbara, CA, USA.

[9] Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. Computer Speech and Language , 14(4):373–400.

[10] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. Numerical Recipes in C: The Art of Scientific Computing . Cambridge University Press.