The LIG English to French Machine Translation System for IWSLT 2012

Laurent Besacier, Benjamin Lecouteux, Marwen Azouzi, Luong Ngoc Quang

LIG
University of Grenoble
firstname.lastname@imag.fr

Abstract

This paper presents the LIG participation to the E-F MT task of IWSLT 2012. The primary system proposed made a large improvement (more than 3 point of BLEU on tst2010 set) compared to our last year participation. Part of this improvment was due to the use of an extraction from the Gigaword corpus. We also propose a preliminary adaptation of the driven decoding concept for machine translation. This method allows an efficient combination of machine translation systems, by rescoring the log-linear model at the N-best list level according to auxiliary systems: the basis technique is essentially guiding the search using one or previous system outputs. The results show that the approach allows a significant improvement in BLEU score using Google translate to guide our own SMT system. We also try to use a confidence measure as an additional log-linear feature but we could not get any improvment with this technique.

1. Introduction

This paper describes LIG approach for the evaluation campaign of the 2012 International Workshop on Spoken Language Translation (IWSLT-2012), English-French MT task. This year the LIG participated only to the E-F MT task and focused on the use of driven decoding to improve statistical machine translation. In addition, we used much more parallel data than last year (trying to make use of the Giga-10^9 corpus). Some (un-successful) attempts to use confidence measures to re-rank our N-best hypotheses were also investigated. The remainder of the paper is structured as follows. Section 2 describes the data we used for training our translation and language models. Section 3 presents the concept of driven decoding that allowed us to get improvements using an auxiliary translation (of an online system) to guide the decoding process. Section 4 presents our attempt to use confidence measures and section 5 details the experiments as well as the LIG official results obtained this year.

2. Resources used in 2012

The following sections describe the resources used to build the translation models as well as the language models.

2.1. Translation models training data

We built three translation models for our machine translation systems (see table 1).

- An in-domain translation model trained on TED Talks collection (TED) corpus.
- A (bigger) out-of-domain translation model trained on six different (freely available) corpora in which three of them are part of the WMT 2012 shared task training data:
  - the latest version of the Europarl (version 7) corpus (EUROPARL[1] [1])
  - the latest version of the News-Commentary (version 7) corpus (NEWS-C)
  - the United Nations corpus (UN [2] [2])
- We also used the Corpus of Parallel Patent Applications (PCT[3]), the DGT Multilingual Translation Memory of the Acquis Communautaire (DGT-TM [3]), and the EUconst corpus (EU-CONST [4]). These three corpora are all freely available.
- An additional out-of-domain translation model was trained on a subset of the French-English Gigaword corpus (GIGA-5M). After cleaning, the whole Gigaword corpus was sorted at sentence level according to the sum of perplexities of the source (English) and the target (French) based on two French and English pretrained language models. For this, LMs were trained separately on all the data listed in table 2 except the Gigaword corpus itself (the News Shuffle corpus was also available on the source English side). The separate LMs were then interpolated using weights estimated on dev2010 using EM algorithm (more details on this process are given in the next section). Finally, the GIGA-5M subset was obtained after filtering out the whole Gigaword corpus with a cut-off limit of 300 (ppl). This leads to a subset of 5M aligned sentences.

[1] http://www.statmt.org/europarl/
[2] http://www.euromatrixplus.net/multi-un/
[3] http://www.wipo.int/patentscope/en/data/pdf/wipo-coppa-technicalDocumentation.pdf

System	Corpus	Aligned Sentences
IN-DOMAIN	TED	139,763
OUT-OF-DOMAIN	EU-CONST	4,904
	NEWS-C	124,081
	EUROPARL	1,743,110
	DGT-TM	1,657,662
	PCT	7,739,299
	UN	10,573,628
Additional GIGA-5M	GIGA-TOP-5M	4,392,530

Table 1: Data used for training the translation model.

Corpus	French words	Alpha	Perplexity
TED	2,798,705	0.536023	103.5
EU-CONST	104,698	5.84281e-06	1074.2
NEWS-C	3,224,063	0.0539594	179.4
EUROPARL	44,116,533	0.119409	156.2
DGT-TM	27,582,544	0.0422644	452.5
PCT	164,936,865	0.0484619	625.3
UN	252,849,705	0.0225498	229.4
NEWS-SHUFFLE	608,297,082	0.0834454	162.2
GIGA-5M	117,985,209	0.131878	141.4

Table 2: Data used for training the language model.

These data were used to train three different translation tables in a multiple phrase table decoding framework (corresponding to the *either* option defined in the Moses advanced features).

2.2. Language model training data

For the language model training, in addition to the French side of all of the parallel corpora described above, we used the News Shuffle corpus provided by the WMT 2012 shared task. First a 5-gram back-off interpolated language model with the modified (improved) Kneser-Ney smoothing was trained on each resource using the SRI language modeling toolkit [5]. Then we created a merged LM optimized on a development corpus (dev2010) using EM algorithm. The details on these LM resources and their weights are given in table 2. The table shows that the in-domain data obviously have a strong weight and that the LM trained on Gigaword subset is also well matched to the TED task. On the contrary, the 3 additional corpora PCT, DGT-TM and EU-CONST are the ones that lead to the highest perplexities and they seem quite far from the TED domain (PCT covers different topics like patents, EU-CONST is too small and DGT-TM covers a topic too far from TED).

2.3. Development and test sets

The TED dev2010 set (934 aligned sentences) was used for tuning and the TED tst2010 set (1 664 aligned sentences) was used for testing and making a choice on the best systems to be presented at the evaluation. These sets will be referred to as dev2010 and tst2010 in the rest of this paper. In addition, the TED tst2011 set (818 aligned sentences) and the TED tst2012 set (1 124 aligned sentences) were used for the official evaluation.

2.4. Data pre-processing

This year we used a fully in-house pre-processing. The goal was to use a more specific pre-processing and post-processing steps for English as well as for French. In short, we applied the following steps:

- filter out badly aligned sentences (using several heuristics)
- filter out empty sentences and sentences having more than 50 words
- filter out pairs of sentences where the ratio is more than 9
- punctuation normalization (extra punctuation mark deletion, transform several encodings of a same punctuation mark function to a canonical version, etc.)
- tokenize (different to the default Moses tokenizer using French grammar rules)

- truecase (remove case for the words at the beginning of the sentence while keeping information on the word position)
- spell correction on both source and target sides
- diacritics restoration (notably on uppercase letters at the beginning of sentences)
- Unicode normalization (NFKC)
- normalization of several words (e.g. coeur)
- disambiguate abbreviations and clitics
- HTML entities conversion

To clean the GigaWord corpus, we applied additional cleaning steps. Many heuristics (rules) were used in order to keep only good quality bi-texts.

2.5. System configuration

In the experiments reported here, 26 or 38 features (according to the total number of PT used) were used in our statistical machine translation system: 10 or 15 translation model scores, 14 or 21 distortion scores, 1 LM score, and 1 word penalty score. We used the Minimum Error Rate Training (MERT) method to tune the weights on dev2010 corpus. We are aware that in the future better optimization techniques like MIRA should be used for such a large number of parameters.

3. Driven Decoding for SMT

Recently, the concept of driven decoding (DD), introduced by [6] has been successfully applied to the automatic speech recognition (speech-to-text) task. This idea is to use an auxiliary transcription (coming from another system output or from another source of information) to guide the decoding process. There is a strong interest in applying this concept to statistical machine translation (SMT). The potential applications are: system combination, multi-source translation (from several languages, from several ASR outputs in the case of speech translation), use of an online system (like Google-translate) as auxiliary translation, on-line hypothesis re-calculation in a post-edition interface, etc.

In short, our first attempt in driven decoding consists in adding several feature functions corresponding to the distance between the current hypothesis decoded (called H) and the auxiliary translation available (T) : d(T,H). Different estimation methods to calculate d(T,H) can be proposed : edit-distance, metrics based-on information theory (entropy, perplexity), metrics based on n-gram coverage (BLEU), etc. As a first attempt, we started to experiment in a re-scoring framework for which N-Best hypotheses from the baseline MT system are re-ordered after adding the new feature functions proposed.

3.1. Related Work

This section presents a brief description of related works. They are found mainly in system combination for both speech recognition and machine translation. Unlike speech recognition, system combination in statistical machine translation involves systems based on potentially different standards such as phrasal, hierarchical and syntax based. This introduces new issues such as breaking up of phrases and alterations of word order. We first propose a description of the application of Driven Decoding (DD) algorithm in ASR systems. Then, various system combination attempts in Machine Translation are presented. Detailed presentation of these two concepts - DD and SMT systems combination - is needed to understand our approach.

3.1.1. Imperfect transcript driven speech recognition

In the paper introduced by [6], the authors try to make use of auxiliary textual information associated with speech signals (such as subtitles associated to the audio channel of a video) to improve speech recognition performance. It is demonstrated that those imperfect transcripts which result in misalignments between the speech and text could actually be taken advantage of. In brief, two methods were proposed. The first method involved the combination of generic language model and a language model estimated on the imperfect transcript resulting in cutting down the linguistic space. The second method involved modifying the decoding algorithm by rescoring the estimate function. The probability of the current hypothesis which results from partial exploration of the search graph is dynamically rescored based on the alignment (with imperfect transcript) scores (done using Dynamic Time Warping). The experimental results which used both dynamic synchronization and linguistic rescoring displayed interesting gains. Another kind of imperfect transcript that can be used is the output hypothesis of another system, leading to an integrated approach for system combination. Thus, in the same paper is proposed a method in which the outputs of the contrastive system drives the decoder of the primary system. The results showed that the new system run by driven decoding algorithm outperformed both primary and contrastive systems. Various cross adaptation schemes were also examined. The principle proposed is that firstly, one-best hypothesis is generated from the auxiliary system and a confidence score is evaluated for each word. Then these informations are used to dynamically modify the linguistic score during decoding. The method was evaluated on a radio broadcast transcription task and it was found that WER reduced significantly (about 1.9%) . The WER gain was even better (2.9%) by combining DD and cross adaptation.

3.1.2. System Combination for Machine Translation

-Confusion Network (CN) Decoding

There are important issues to address for machine translation system combination using confusion network decoding. An important one is the presence of errors in the alignment of hypotheses which lead to ungrammatical combination outputs. [7] proposed arbitrary features that can be added log-linearly into the objective function in this method. This addition of new features is the core idea we followed in our proposal.

Confusion Network decoding for MT system combination has been proposed in [8]. The hypothesis have to be aligned using Levenshtein alignment to generate the confusion network. One hypothesis is chosen as skeletal hypothesis and others are aligned against it. In [7], 1-best output from each system is used as the skeleton to develop the confusion network and the average of the TER scores between the skeleton and other hypotheses were used to evaluate the prior probability. Finally a joint lattice is generated by aggregating all the confusion networks parallely. Through this work it is shown that arbitrary features could be added log-linearly by evaluating log-posterior probabilities for each confusing network arc. In confusion network decoding, the word order of the combination is affected by the skeletal hypothesis. Hence the quality of the output from the combination also depends on the skeletal hypothesis. The hypothesis with the minimum average TER-score on aligning with all other hypothesis is proposed as an improved skeletal hypothesis.

$$E_s = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} TER(E_j, E_i) \qquad (1)$$

where N_s is the number of systems and E_s is the skeletal hypothesis.

In [9] system specific confidence scores are also introduced. The better the confidence score the higher the impact of that system. In the experimental part of this same work, three phrase-based (A,C,E), two hierarchical (B,D) and one syntax based (F) systems are combined. All of them are trained on the same data. The decoder weights are tuned to optimize TER for systems A and B and BLEU for the remaining systems. Decoder weight tuning is done on the NIST MT02 task. The results of the combination system were better than single system on all the metrics but for only TER and BLEU tuning. In the case of METEOR tuning, the combination system produced high TER and low BLEU score. The experiments were performed on Arabic and Chinese NIST MT tasks.

-N-Best Concatenation and Rescoring

Another paper [10] presents a slightly different method where N-Best hypotheses are re-scored instead of building a synthesis (CN) of the MT outputs (as described in previous sub-section). The N-Best list from all input systems are combined and then the best hypothesis is selected according to feature scores. Three types of features are: language model features, lexical features, N-Best list based features. The feature weights are modified using Minimum Error Rate Training (MERT). Experiments are performed to find the optimal size for N-Best list combination. Four systems are used and analysed on combination of two best systems and all the systems. 50-best list was found to be optimal size for both cases. The authors showed that the impact of gradually introducing a new system for combination becomes lower as the number of systems increases. Anyway the best result is obtained when all of the systems are combined.

-Co-decoding

Recently, the concept of collaborative decoding (co-decoding) was introduced by [11] to improve machine translation accuracy by leveraging translation consensus between multiple machine translation decoders. Different from what we described earlier (postprocess the n-best lists or word graphs), this method uses multiple machine translation decoders that collaborate by exchanging partial translation results. Using an iterative decoding approach, n-gram agreement statistics between translations of multiple decoders are employed to re-rank full and partial hypotheses explored in decoding.

3.2. Overview of the Driven Decoding Concept

3.2.1. Driven Decoding

As said in the introduction part, driven decoding consists in adding several feature functions to the log-linear model before N-Best list re-ordering. Practically, after N-Best lists are generated by an individual system, additional scores are added to each line of the N-Best list file. Theses additional scores correspond to the distance between the current hypothesis decoded (called H) and the auxiliary translation available (T) : d(T,H). Let's say that 2 auxiliary translations are available (from system 1 and system 2) and that 4 distance metrics are available (BLEU, TER, TERp-A and PER); in that case, 8 scores are added to each line of the N-Best list. The distance metrics used in our experiments are described in the next section and then N-Best reordering process is detailed.

3.2.2. Distance Metrics used

The distance metrics used are Translation Error Rate (TER), Position independent Error Rate (PER), TERp-A and BLEU [12]. The TER score reflects the number of edit operations (insertions, deletions, words substitutions and blocks shifts) needed to transform a hypothesis translation into the reference translation, while the BLEU score is the geometric mean of n-gram precision. Lower TER and higher BLEU score suggest better translation quality. In addition, we use PER score (position independent error rate) which can be seen as a bag-of-words metric potentially interesting in the context of the driven decoding proposed. In addition we use TERp [13] which is an extension of TER eliminating its

shortcomings by taking into account the linguistic edit operations, such as stem matches, synonyms matches and phrase substitutions besides the TER's conventional ones. These additions allow us to avoid categorizing the hypothesis word as Insertion or Substitution in case that it shares same stem, or belongs to the same synonym set represented by WordNet, or is the paraphrase of word in the reference. More precisely, we used TERp-A, another version of TERp, in which each above mentioned edit cost has been tuned to maximize the correlation with human judgment of Adequacy at the segment level (from the NIST Metrics MATR 2008 Challenge development data). However, it is worth mentionning that for this particular task, we use a degraded version of TERp-A which does not take into account synonymy, because the target language is French while the TERp-A metric only implements the use of (English) Wordnet.

3.2.3. N-Best Reordering and Combination

In this framework the system combination is based on the 1000-best outputs (we generally have less on IWSLT data) generated by the LIG primary system using the "uniq" option. Our primary system uses 3 different translation and re-ordering tables. So each N-best list is associated with a set of 38 scores: 1 LM score, 15 translation model scores, 1 distance-based reordering score, 21 lexicalized reordering scores. In addition we introduce 8 distance metrics scores for each sentence.

-The training step

The score combination weights are optimized in order to maximize the BLEU score at the sentence level. This step is performed by using the MERT tool. The weights of "standard" scores are initialized with the tuned weights computed during the usual tuning phase. In a second time, we fine tune weights of the introduced distance metrics (this can be seen as an additional iteration of MERT).

-The decoding step

The decoding step combines all the scores: a global score is computed for each sentence (i.e. the log-linear score) and sentences are reordered according to the final combined score.

4. Use of Confidence Measures for SMT

Besides driven decoding (DD) scores, a sentence confidence score can be added as an additional feature in the N-best list to improve the re-ordering performance. To obtain such a confidence score, a classifier must be constructed. We concatenate two data sets dev2010 + tst2010 to form the training data. Features used to train our model come from the baseline features of the WMT2012 quality estimation shared task (features originally presented in [14]), which can be summarized as follows:

- Source and target sentence: number of tokens and their ratio, number of punctuation marks.
- Source and target sentence's language model probabilities.
- Percentage of unigrams / bigrams / trigrams in quartiles 1 (and 4) of frequency (lower and higher frequency ngrams) in a corpus of the source language.
- Average number of translation per source word in the sentence, unweighted or weighted by the inverse frequency of each word in the source corpus.

The core element needed for the classifier construction process is the training label for each sentence. The TERp-A metric [13], which we select to perform this task, provides the linguistic and semantic matching between each sentence in training set and its reference (available for dev2010 and tst2010 corpora), then yields the minimum cost for matching normalized by its number of tokens as its score. We then categorize them in a binary set: sentences with score higher than 0.3 is assigned with "Good" (G) label, otherwise, "Bad" (B). A CRF-based toolkit, WAPITI [15], is then called to build the classifier. The training phase is conducted using stochastic gradient descent (SGD-L1) algorithm, with values for maximum number of iterations done by the algorithm (-maxiter), stop window size (–stopwin) and stop epsilon (–stopeps) to 200, 6, and 0.00005 respectively.
Applying this classifier in both test sets (test2011 + test2012, with WAPITI's default threshold = 0.5) gives us the result files detailing hypothesized label along with its probability at the sentence level. Then, the confidence score used is the probability of sentence to be regarded as a "Good" sentence. For instance, a sentence classified as "G" with related probability of 0.8 gets obviously the confidence score of 0.8; meanwhile the other one labeled as "B" with probability of 0.7 will have the score of 0.3. This score is used as an additional feature in the log-linear model just as it is done for driven decoding (see previous section).
Performance of the re-ordering task with and without the use of confidence measure will be shown in Table 3.

5. Experimental Results of LIG Systems

We recall that our systems were systematically tuned on dev2010 corpus. Our baseline system, trained as described in section 2, lead to a BLEU score of 30.28 on tst2010 using 2 translation and re-ordering models (no GIGAword) while it improves to 30.80 using 3 translation and reordering models (using GIGAword). This result has to be compared with 27.58 obtained on tst2010 with our system last year.

As far as the driven decoding is concerned, the results show that using the Google 1best hypothesis to guide the

system	dev2010	tst2010	tst2011	tst2012	submission
Baseline (2TM)	27.41	30.28	x	x	
Baseline+GIGAword (3TM)	27.84	30.80	36.88	37.58	**primary**
+DD-google	28.69	**32.01**	39.09	39.36	**contrastive**
+conf	27.84	30.80	x	x	
+DD-google+conf	**28.77**	31.87	x	x	
+DD-ref	32.84	37.26	x	x	oracle
online-google	26.90	33.77	40.16	x	

Table 3: Performances (BLEU case+punct) for several LIG systems

rescoring of the LIG Nbest list leads to significant improvements on all data sets. On dev2010 data, the performance obtained is even better that both LIG and Google systems evaluated separately. On tst2010 and tst2011 the driven decoding is slightly below google. This can be explained by the fact that google has a very different behavior from one set to another (on the dev google is significantly worse than LIG system while he gets better results on tst2011). The LIG system driven by Google 1best was, however, not submitted as a primary run since we used an online system to improve our own module (contrastive system).

On the contrary, adding confidence measures gives only slight improvement on the dev2010 set and does not generalize on tst2010 so it was finally not used in our final submission. According to our analysis, this unsuccessful experiment can be originated from the following reasons: (1) The feature set is simply and superficially constructed hence fails to cover all aspect of quality. This hypothesis can motivate us to explore more types of features (lexical, syntactic, semantic...) in the future work ; (2) the whole combination of features without any selection strategy might be an unskilful option weakening our classifier capability. For information, the oracle obtained, using the golden reference as an auxiliary system, is given in the last line of the table, as well as the performance of the online Google system.

6. Conclusions

This paper described the LIG participation to the E-F MT task of IWSLT 2012. The primary system proposed made a large improvement (more than 3 point of BLEU on tst2010 set) compared to our last year participation. Part of this improvement was due to the use of an extraction from the Gigaword corpus. We have proposed a preliminary adaptation of the driven decoding concept for machine translation. This method allows an efficient combination of machine translation systems, by rescoring the log-linear model at the N-best list level according to auxiliary systems: the basis technique is essentially guiding the search using one or previous system outputs. The results show that the approach allows a significant improvement in BLEU score using Google translate to guide our own SMT system (such system was submitted as contrastive since it uses an online translation). We also tried to use a confidence measure as an additional log-linear feature but we could not get any improvement with this technique.

7. References

[1] P. Koehn, "Europarl: a parallel corpus for statistical machine translation." in *MT Summit X*, Phuket, Thailand, 2005, pp. 79–86.

[2] A. Eisele and Y. Chen, "Multiun: a multilingual corpus from united nation documents." in *LREC 2010*, Valletta, Malta, 2010, pp. 2868–287.

[3] R. Steinberger, A. Eisele, S. Klocek, P. Spyridon, and P. Schlter, "Dgt-tm: A freely available translation memory in 22 languages," in *LREC 2012*, Istanbul, Turkey, 2012.

[4] J. Tiedemann, "News from opus - a collection of multilingual parallel corpora with tools and interfaces," in *Recent Advances in Natural Language Processing*, 2009.

[5] A. Stolcke, "SRILM — an extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA, 2002.

[6] B. Lecouteux, G. Linares, and S. Oger, "Integrating imperfect transcripts into speech recognition systems for building high-quality corpora," *Computer Speech and Language*, vol. 26, no. 2, pp. 67 – 89, 2012.

[7] A.-v. Rosti, S. Matsoukas, and R. Schwartz, "Improved word-level system combination for machine translation," in *In Proceedings of ACL*, 2007.

[8] S. Bangalore, "Computing consensus translation from multiple machine translation systems," in *In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001*, 2001, pp. 351–354.

[9] A.-v. Rosti, N.-F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz, and B. Dorr, "Combining outputs from multiple machine translation systems," in *In Proceedings of the North American Chapter of the Association*

for Computational Linguistics Human Language Technologies, 2007, pp. 228–235.

[10] A. S. Hildebrand and S. Vogel, "Combination of machine translation systems via hypothesis selection from combined n-best lists," in *Proceedings of Association for Machine Translation in the Americas (AMTA)*, Hawai, USA, 2009.

[11] M. Li, N. Duan, D. Zhang, C.-h. Li, and M. Zhou, "Collaborative decoding: Partial hypothesis re-ranking using translation consensus between decoders," in *Joint ACL IJCNLP*, 2009.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation." in *ACL*. ACL, 2002.

[13] M. Snover, N. Madnani, B. Dorr, and R. Schwartz, "Terp: A system description," in *Proceedings of the First NIST Metrics for Machine Translation Challenge (MetricsMATR)*, Waikiki, Hawaii, October 2008.

[14] L. Specia, M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini, "Estimating the sentence-level quality of machine translation systems," in *13th Conference of the European Association for Machine Translation*, Barcelona, Spain, 2009, p. 2837.

[15] T. Lavergne, O. Cappe, and F. Yvon, "Practical very large scale crfs," in *Proceedings ACL*, 2010, p. 504513.