The LIG machine translation system for WMT 2010

Marion Potet, Laurent Besacier and Hervé Blanchon
LIG Laboratory, GETALP Team
University Joseph Fourier, Grenoble, France.
Marion.Potet@imag.fr
Laurent.Besacier@imag.fr
Herve.Blanchon@imag.fr

Abstract

This paper describes the system submitted by the Laboratory of Informatics of Grenoble (LIG) for the fifth Workshop on Statistical Machine Translation. We participated to the news shared translation task for the French-English language pair. We investigated differents techniques to simply deal with Out-Of-Vocabulary words in a statistical phrase-based machine translation system and analyze their impact on translation quality. The final submission is a combination between a standard phrase-based system using the Moses decoder, with appropriate setups and pre-processing, and a lemmatized system to deal with Out-Of-Vocabulary conjugated verbs.

1 Introduction

We participated, for the first time, to the shared news translation task of the fifth Workshop on Machine Translation (WMT 2010) for the French-English language pair. The submission was performed using a standard phrase-based translation system with appropriate setups and pre-processings in order to deal with system's unknown words. Indeed, as shown in (Carpuat, 2009), (Habash, 2008) and (Niessen, 2004), handling Ou-of-Vocabulary words with techniques like lemmatization, phrase table extension or morphological pre-processing is a way to improve translation quality. After a short presentation of our baseline system setups we discuss the effect of Out-Of-Vocabulary words in the system and introduce some ideas we chose to implement. In the last part, we evaluate their impact on translation quality using automatic and human evaluations.

2 Baseline System Setup

2.1 Used Resources

We used the provided Europarl and News parallel corpora (total 1,638,440 sentences) to train the translation model and the News monolingual corpora (48,653,884 sentences) to train the language model. The 2008 News test corpora (news-test2008; 2,028 sentences) was used to tune the produced system and last year's test corpora (news-test2009; 3,027 sentences) was used for evaluation purposes. These corpora will be refered to as *Dev* and *Test* later in the paper. As preprocessing steps, we applied the PERL scripts provided with the corpora to lowercase and tokenise the data.

2.2 Language modeling

The target language model is a standard n-gram language model trained using the SRI language modeling toolkit (Stocke, 2002) on the news monolingual corpus. The smoothing technique we applied is the modified Kneser-Ney discounting with interpolation.

2.3 Translation modeling

The translation model was trained using the parallel corpus described earlier (Europarl+News). First, the 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 (Koehn et al., 2007). 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 our experiment we used thirteen standard translation models: six distortion models, a lexicon word-based and a phrase-based translation model for both direction, and a phrase, word and distortion penalty.

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 167–172,
Uppsala, Sweden, 11-16 July 2010.

2.4 Tuning and decoding

For the decoding (i.e. translation of the test set), the system uses a log-linear combination of the previous target language model and the thirteen translation models extracted from the phrase-table. As the system can be beforehand tuned by adjusting log-linear combination weights on a developement corpus, we used the Minimum Error Rate Training (MERT) method, by (Och, 2003).

3 Ways of Improvements

3.1 Discussion about Out-Of-Vocabulary words in PBMT systems

Phrase-based statistical machine translation (PBMT) use phrases as units in the translation process. A phrase is a sequence of n consecutive words known by the system. During the training, these phrases are automaticaly learned and each source phrase is mapped with its corresponding target phrase. Throughout test set decoding, a word not being part of this vocabulary list is labeled as "Out-Of-Vocabulary" (OOV) and, as it doesn't appear in the translation table, the system is unable to translate it. During the decoding, Out-Of-Vocabulary words lead to "broken" phrases and degrade translation quality. For these reasons, we present some techniques to handle Out-Of-Vocabulary words in a PBMT system and combine these techniques before evaluating them.

In a preliminary study, we automatically extracted and manually analyzed OOVs of a 1000 sentences sample extracted from the test corpus (news-test2009). There were altogether 487 OOVs tokens wich include 64.34% proper nouns and words in foreign languages, 17.62% common nouns, 15.16% conjugated verbs, 1.84% errors in source corpus and 1.02% numbers. Note that, as our system is configured to copy systematically the OOVs in the produced translated sentence, the rewriting of proper nouns and words in foreign language is straightforward in that case. However, we still have to deal with common nouns and conjugated verbs.

Initial sentence:
"Cela ne marchera pas" *souligna-t-il* par la suite.
Normalised sentence:
"Cela ne marchera pas" *il souligna* par la suite

Figure 1: Normalisation of the euphonious "t"

3.2 Term expansion with dictionary

The first idea is to expand the vocabulary size, more specifically minimizing Out-Of-Vocabulary common nouns adding a French-English dictionary during the training process. In our experiment, we used a free dictionnary made available by the *Wiktionary*[1] collaborative project (wich aims to produce free-content multilingual dictionaries). The provided dictionnary, containing 15,200 entries, is added to the bilingual training corpus before phrase-table extraction.

3.3 Lemmatization of the French source verbs

To avoid Out-Of-Vocabulary conjugated verbs one idea is to lemmatize verbs in the source training and test corpus to train a so-called lemmatized system. We used the freely available French lemmatiser LIA_TAGG (Béchet, 2001). But, applying lemmatization leads to a loss of information (tense, person, number) which may affect deeply the translation quality. Thus, we decided to use the lematized system only when OOV verbs are present in the source sentence to be translated. Consequently, we differentiate two kinds of sentences: -sentences containing at least one OOV conjugated verb, and - sentences which do not have any conjugated verb (these latter sentences obviously don't need any lemmatization!). Thereby, we decided to build a combined translation system which call the lemmatized system only when the source sentence contains at least one Out-Of-Vocabulary conjugated verb (otherwise, the sentence will be translated by the standard system). To detect sentences with Out-Of-Vocabulary conjugated verb we translate each sentence with both systems (lemmatized and standard), count OOV and use the lemmatized translation only if it contains less OOV than the standard translation. For example, a translation containing k Out-Of-Vocabulary conjugated verbs and n others Out-Of-Vocabulary words (in total $k+n$ OOV) with the standard system, contains, most probably, only n Out-Of-Vocabulary words with the lemmatised system because the conjugated verbs will be lemmatized, recognized and translated by the system.

[1] http://wiki.webz.cz/dict/

3.4 Normalization of a special French form

We observed, in the French source corpra, a special French form which generates almost always Out-Of-Vocabulary words in the English translation. The special French form, named euphonious "t", consists of adding the letter "t" between a verb (ended by "a", "e" or "c") and a personal pronoun and, then, inverse them in order to facilitate the prononciation. The sequence is represented by: *verb*-t-*pronoun* like *annonca-t-elle*, *arrive-t-il*, *a-t-on*, etc. This form concerns 1.75% of the French sentences in the test corpus whereas these account for 0.66% and 0.78% respetively in the training and the developement corpora. The normalized proposed form, illustrated below in figure 1, contains the subject pronoun (in first posistion) and the verb (in the second position). This change has no influence on the French source sentence and accordingly on the correctness and fluency of the English translation.

3.5 Adaptation of the language model

Finally, for each system, we decided to apply different language models and to look at those who perfom well. In addition to the 5-gram language model, we trained and tested 3-gram and 4-gram language models with two different kinds of vocabularies : - the first one (conventional, refered to as n-gram in table 3) contains an open-vocabulary extracted from the monolingual English training data, and - the second one (refered to as n-gram-vocab in table 3) contains a closed-vocabulary extracted from the English part of the bilingual training data. In both cases, language model probabilities are trained from the monolingual LM training data but, in the second case, the lexicon is restricted to the one of the phrase-table.

4 Experimental results

In the automatic evaluation, the reported evaluation metric is the BLEU score (Papineni et al., 2002) computed by MTEval version 13a. The results are reported in table 1. Note that in our experiments, according to the resampling method of (Koehn, 2004), there are significative variations (improvement or deterioration), with 95% certainty, only if the difference between two BLEU scores represent, at least, 0.33 points. To complete this automatic evaluation, we performed a human analysis of the systems outputs.

4.1 Standard systems

4.1.1 Term expansion with dictionary

Regarding the results of automatic evaluation (table 1, system (2)), adding the dictionary do not leads to a significant improvement. The OOV rate and system perplexity are reduced but, ignoring the tuned system which presents lower performance, the BLEU score decreases significatly on the test set. The BLEU score of the system augmented with the dictionary is 24.50 whereas the baseline one is 24.94. So we can conclude that there is not a meaningfull positive contribution, probably because the size of the dictionary is very small regarding the bilingual training corpus. We found out very few Out-Of-Vocabulary words of the standard system recognized by the system with the dictionary, see figure 2 for example (among them : *coupon*, *cafard*, *blonde*, *retardataire*, *médicaments*, *pamplemousse*, etc.). But, as the dictionnary is very small, most OOV common words like *hôtesse* and *clignotant* are still unknown. Regarding the output sentences, we note that there are very few differences and the quality is equivalent. The dictionary used is to small to extend the system's vocabulary and most of words still Out-Of-Vocabulary are conjugated verbs and unrecognized forms.

Baseline system:
A *cafard* fled before the danger, but if he felt fear?
System with dictionary:
A *blues* fled before the danger, but if he felt fear?

Figure 2: Example of sentence with an OOV common noun

4.1.2 Normalisation of special French form

Considering the BLEU score, the normalization of French euphonious "t" have, apparently, very few repercussion on the translation result (table 1, system (3)) but the human analysis indicates that, in our context, the normalisation of euphonious "t" brings a clear improvement as seen in example 3. Consequently, this preprocessing is kept in the final system.

4.1.3 Tuning

We can see in table 1 that the usual tuning with Minimum Error Rate Training algorithm deteriorates systematically performance scores on the test set, for all systems. This can be explained by the

System	OOVs	ppl	Dev score	Test score
(1) Baseline	2.32%	207	29.72 (19.93)	23.77 (24.94)
(2) + dictionary	2.30%	204	30.01 (23.92)	24.32 (24.50)
(3) + normalization	2.31%	204	30.07 (19.90)	23.99 (24.98)
(4) + normalization + Dev data	2.30%	204	/ (/)	/ (25,05)

Table 1: Standard systems BLEU scores with tuning (without tuning)/ LM 5-gram

Baseline system:
"It will not work" *souligna-t-il* afterwards.
System with normalisation:
"It will not work" *he stressed* afterwards.

Figure 3: Example of sentence with a "*verb*-t-*pronoun*" form

gap between the developement and test corpora (ie the Dev set may be not representative of the Test set). So, even if it is recommanded in the standard process, we do not tune our system (we use the default weights proposed by the Moses decoder) and add the developement corpus to train it. In this case, the training set contains 1,640,468 sentences (the initial 1,638,440 sentences and the 2,028 sentences of the developement set). This slightly improves the system (from 24.98, the BLEU score raise to 25,05 after adding the developpement set to the training).

4.2 Lemmatised systems

Results of lemmatised systems are reported on table 2. First, we can notice that, in this particular case, the tuning (with MERT method) is mandatory to adapt the weights of the log linear model. Our analysis of the tuned weight of the lemmatised system shows that, in particular, the word penalty model has a very low weight (this favours short sentences) and the lexical word-based translation models have a very low weight (no use of the lexical translation probability). We also notice that the lemmatization leads to a real drop-off of OOV rate (fall from 2.32% for the baseline, to 2.23% for the lemmatized system) and perplexity (fall from 207 for the baseline, to 178 for the lemmatized system). We can observe a clear decrease of the performance with the lemmatized system (BLEU score of 20.50) compared with a non-lemmatized one (BLEU score of 24.94). This can be significatively improved applying euphonious "t" normalization to the source data (BLEU score of 22.14). Almost all French OOV conjugated verbs with the standard system were recognized by the lemmatized one (*trierait*, *joues*, *testaient*, *immergée*, *économiseraient*, *baisserait*, *prépares*, etc.) but the small decrease of the translation quality can be explained, among other things, by several tense errors. See illustration in figure 4. So, we conclude that the systematic normalization of French verbs, as a pre-process, reduce the Out-Of-Vocabulary conjugated verbs but decrease slighly the final translation quality. The use of such a system is helpfull especially when the sentence contains conjugated verbs (see example 5).

4.3 Adaptation of the language model

We applied five differents language models (3-gram and 4-gram language models with selected vocabulary or not and a 5-gram language model) to the four standard systems and the two lemmatised one. The results, reported in table 3, show that BLEU score can be significantly different depending on the language model used. For example, the fifth system (5) obtained a BLEU score of 21.48 with a 3-gram language model and a BLEU score of 22.84 with a 4-gram language model. We can also notice that five out of our six systems outperform using a language model with selected vocabulary (*n-gram-vocab*). One possible explanation is that with LM using selected vocabulary (*n-gram-vocab*), there is no loss of probability mass for english words not present in the translation table.

4.4 Final combined system

Considering the previous observations, we believe that the best choice is to apply the lemmatized system only if necessary i.e. only if the sentence contains OOV conjugated verbs, otherwise, a standard system should be used. We consider system (4), with 4-gram-vocab language model (selected vocabulary) without tuning, as the best standard system and system (6), with 3-gram-vocab language model (selected vocabulary) not tuned either, as the best lemmatized system. The final

System	OOVs	ppl	Dev score	Test score
(5) lemmatization	2.23%	178	20.97 (8.57)	20.50 (8.56)
(6) lemmatization + normalization	2.18%	175	27.81 (9.20)	22.14 (10.82)

Table 2: Lemmatised systems BLEU scores with tuning (without tuning)/ LM 5-gram

Baseline system: You *will be limited* by the absence of exit for headphones.
Lemmatised system: You *are limited* by the lack of exit for ordinary headphones.
reference: You *will be limited* by the absence of output on ordinary headphones.

Figure 4: Example of sentences without OOV verbs

system translations are those of the lemmatized system (6) when we translate sentences with one or more Out-Of-Vocabulary conjugated verbs and those of the un-lemmatized system (4) otherwise. Around 6% of test set sentences were translated by the lemmatized system. Considering the results reported in table 4, the combined system's BLEU score is comparable to the standard one (25.11 against 25.17).

System	Test score	sentences
(4) Standard sys.	25.17	94 %
(6) Lemmatised sys.	22.89	6%
(7) Combined	25.11	100 %

Table 4: Combined system's results and % translated sentences by each system

5 Human evaluation

We compared two data set. The first set (selected sent.) contains 301 sentences selected from test data by the combined system (7) to be translated by the lemmatized system (6) whereas the second set (random sent.) contains 301 sentences randomly picked up. The latter is our control data set. We compared for both groups the translation hypothesis given by the lemmatized system and the standard one.

We performed a subjective evaluation with the NIST five points scales to measure fluency and adequacy of each sentences through SECtra_w interface (Huynh et al., 2009).We involved a total of 6 volunteers judges (3 for each set). We evaluated the inter-annotator agreement using a generalized version of Kappa. The results show a *slight* to *fair* agreement according (Landis, 1977).

The evaluation results, detailled in table 5 and 6, showed that both fluency and adequacy were improved using our combined system. Indeed, for a random input (random sent.), the lemmatized system lowers the translations quality (fluency and adequacy are degraded for, respectively, 35.8% and 37.5% of the sentences), while it improves the quality for sentences selected by the combined system (for "selected sent.", fluency and adequacy are improved or stable for 81% of the sentences).

Adequacy	selected sent.	random sent.
(6) $\geq$ (4)	81%	62.4%
(6) $<$ (4)	18.9%	37.5%

Table 5: Subjective evaluation of sentences adequacy ((6) lemmatized system - (4) standard system)

Fluency	selected sent.	random sent.
(6) $\geq$ (4)	81%	64.1%
(6)$<$(4)	18.9%	35.8%

Table 6: Subjective evaluation of sentences fluency ((6) lemmatized system - (4) standard system)

6 Conclusion and Discussion

We have described the system used for our submission to the WMT'10 shared translation task for the French-English language pair.

We propose dsome very simple techniques to improve rapidely a statistical machine translation. Those techniques particularly aim at handling Out-Of-Vocabulary words in statistical phrase-based machine translation and lead an improved fluency in translation results. The submited system (see section 4.4) is a combination between a standard system and a lemmatized system with appropriate setup.

Baseline system: At the end of trade, the stock market in the negative *bascula*.
Lemmatised system: At the end of trade, the stock market exchange *stumbled* into the negative.

Baseline system: You can choose *conseillera*.
Lemmatised system: We would *advise* you, how to choose.

Figure 5: Example of sentences with OOV conjugated verbs

System	3-gram	3-gram-vocab	4-gram	4-gram-vocab	5-gram
(1)	24.60	24.95	24.94	**25.11**	24.94
(2)	25.14	**25.17**	24.50	23.49	24.50
(3)	24.88	25.00	24.98	**25.15**	24.98
(4)	24.92	24.99	25.05	**25.17**	25.05
(5)	21.48	19.48	**22.84**	20.18	20.50
(6)	22.60	**22.89**	22.14	22.24	22.14

Table 3: Systems's results on test set with differents language models

This system evaluation showed a positive influence on translation quality, indeed, while the improvements on automatic metrics are small, manual inspection suggests a significant improvements of translation fluency and adequacy.

In future work, we plan to investigate and develop more sophisticated methods to deal with Out-Of-Vocabulary words, still relying on the analyze of our system output. We believe, for example, that an appropriate way to use the dictionary, a sensible pre-processings of French source texts (in particular normalization of some specific French forms) and a factorial lemmatization with the tense information can highly reduce OOV rate and improve translation quality.

References

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*. Prague, Czech Republic.

Papineni K., Roukos S., Ward T., and Zhu W.J. 2002. BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, pp. 311–318. Philadelphia, Pennsylvania, USA.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. *International Conference on Spoken Language Processing*, Vol. 2, pp 901–904. Denver, Colorado, USA.

Frederic Béchet. 2001. LIA_TAGG. http://old.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*. Sapporo, July.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. *conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp 388–395. Barcelona, Spain.

Marine Carpuat. 2009. Toward Using Morphology in French-English Phrase-based SMT. *Workshop on Machine Translation in European Association for Computational Linguistics (EACL-WMT)*, pp 150–154. Athens, Greece.

Sonja Niessen and Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, vol. 30, pp 181–204.

Nizar Habash. 2008. Four techniques for Online Handling of Out-Of-Vocabulary Words in Arabic-English Statistical Machine Translation. *Human Language Technology Workshop in Association for Computational Linguistics, (ACL-HTL)*, pp 57–60. Columbus, Ohio, USA.

Landis J. R. and Koch G. G.. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, vol. 33, pp. 159–174.

Hervé Blanchon, Christian Boitet and Cong-Phap Huynh. 2009. A Web Service Enabling Gradable Post-edition of Pre-translations Produced by Existing Translation Tools: Practical Use to Provide High-quality Translation of an Online Encyclopedia. *MT Summit XII, Beyond Translation Memories: New Tools for Translators Workshop*, pp 20–27. Ottawa, Canada.