LIG System for Word Level QE task at WMT14

Ngoc-Quang Luong **Laurent Besacier** **Benjamin Lecouteux**
LIG, Campus de Grenoble
41, Rue des Mathématiques,
UJF - BP53, F-38041 Grenoble Cedex 9, France
{ngoc-quang.luong,laurent.besacier,benjamin.lecouteux}@imag.fr

Abstract

This paper describes our Word-level QE system for WMT 2014 shared task on Spanish - English pair. Compared to WMT 2013, this year's task is different due to the lack of SMT setting information and additional resources. We report how we overcome this challenge to retain most of the important features which performed well last year in our system. Novel features related to the availability of multiple systems output (new point of this year) are also proposed and experimented along with baseline set. The system is optimized by several ways: tuning the classification threshold, combining with WMT 2013 data, and refining using *Feature Selection* strategy on our development set, before dealing with the test set for submission.

1 Introduction

1.1 Overview of task 2 in WMT14

This year WMT calls for methods which predict the MT output quality at run-time, on both levels: sentence (Task 1) and word (Task 2). Towards a SMT system-independent and widely-applied estimation, MT outputs are collected from multiple translation means (machine and human), therefore all SMT specific settings (and the associated features that could have been extracted from it) become unavailable. This initiative puts more challenges on participants, yet motivates number of SMT-unconventional approaches and inspires the endeavors aiming at an "Evaluation For All".

We focus our effort on Task 2 (Word-level QE), where, unlike in WMT2013, participants are requested to generate prediction labels for words in three variants:

- Binary: words are judged as *Good* (no translation error), or *Bad* (need for editing).
- Level 1: the *Good* class is kept intact, whereas *Bad* one is further divided into subcategories: *Accuracy* issue (the word does not accurately reflect the source text) and *Fluency* issue (the word does not relate to the form or content of the target text).
- Multi-class: more detailed judgement, where the translation errors are further decomposed into 16 labels based on MQM[1] metric.

1.2 Related work

WMT 2013 witnessed several attempts dealing with this evaluation type in its first launch. Han et al. (2013); Luong et al. (2013) employed the Conditional Random Fields (CRF) (Lafferty et al., 2001) model as their Machine Learning method to address the problem as a sequence labeling task. Meanwhile, Bicici (2013) extended the global learning model by dynamic training with adaptive weight updates in the perceptron training algorithm. As far as prediction indicators are concerned, Bicici (2013) proposed seven word feature types and found among them the "common cover links" (the links that point from the leaf node containing this word to other leaf nodes in the same subtree of the syntactic tree) the most outstanding. Han et al. (2013) focused only on various n-gram combinations of target words. Inheriting most of previously-recognized features, Luong et al. (2013) integrated a number of new indicators relying on graph topology, pseudo reference, syntactic behavior (constituent label, distance to the semantic tree root) and polysemy characteristic. Optimization endeavors were also made to enhance the baseline, including classification threshold tuning, feature selection and boosting technique (Luong et al., 2013).

[1] http://www.qt21.eu/launchpad/content/training

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 335–341,
Baltimore, Maryland USA, June 26–27, 2014.

1.3 Paper outline

The rest of our paper is structured as follows: in the next section, we describe 2014 provided data for Task 2, and the additional data used to train the system. Section 3 lists the entire feature set, involving WMT 2013 set as well as a new feature proposed for this year. Baseline system experiments and methods for optimizing it are furthered discussed in Section 4 and Section 5 respectively. Section 6 selects the most outstanding system for submission. The last section summarizes the approach and opens new outlook.

2 Data and Supporting Resources

For English - Spanish language pair in Task 2, the organizers released two bilingual data sets: the training and the test ones. The training set contains 1.957 MT outputs, in which each token is annotated with one appropriate label. In the binary variant, the words are classified into *"OK"* (no translation error) or *"BAD"* (edit operators needed) label. Meanwhile, in the level 1 variant, they belong to *"OK"*, *"Accuracy"* or *"Fluency"* (two latter ones are divided from *"BAD"* label of the first subtask). In the last variant, multi-class, beside *"Accuracy"* and *"Fluency"* we have further 15 labels based on MQM metric: *Terminology, Mistranslation, Omission, Addition, Untranslated, Style/register, Capitalization, Spelling, Punctuation, Typography, Morphology_(word_form), Part_of_speech, Agreement, Word_order, Function_words, Tense/aspect/mood, Grammar* and *Unintelligible*. The test set consists of 382 sentences where all the labels accompanying words are hidden. For optimizing parameters of the classifier, we extract last 200 sentences from the training set to form a development (dev) set. Besides, the Spanish - English corpus provided in WMT 2013 (total of 1087 tuples) is also exploited to enrich our WMT 2014 system. Unfortunately, 2013 data can only help us in the binary variant, due to the discrepancy in training labels. Some statistics about each set can be found in Table 1.

In addition, additional (MT-independent) resources are used for the feature extraction, including:

- Spanish and English Word Language Models (LM)
- Spanish and English POS Language Models
- Spanish - English 2013 MT system

On the contrary, no specific MT setting is provided (e.g. the code to re-run Moses system like WMT 2013), leading to the unavailability of some crucial resources, such as the *N*-best list and alignment information. Coping with this, we firstly thought of using the Moses "Constrained Decoding" option as a method to tie our (already available) decoder's output to the given target translations (this feature is supported by the latest version of Moses (Koehn et al., 2007) in 2013). Our hope was that, by doing so, both *N*-best list and alignment information would be generated during decoding. But the decoder failed to output all translations (only $1/4$ was obtained) when the number of allowed unknown words (*-max-unknowns*) was set as 0. Switching to non zero value for this option did not help either since, even if more outputs were generated, alignment information was biased in that case due to additional/missing words in the obtained MT output. Ultimately, we decided to employ GIZA++ toolkit (Och and Ney, 2003) to obtain at least the alignment information (and associated features) between source text and target MT output. However, no *N*-best list were extracted nor available as in last year system. Nevertheless, we tried to extract some features equivalent to last year N-best features (details can be found in Section 3.2).

3 Feature Extraction

In this section, we briefly list out all the features used in WMT 2013 (Luong et al., 2013) that were kept for this year, followed by some proposed features taking advantage of the provided resources and multiple translation system outputs (for a same source sentence).

3.1 WMT13 features

- Source word features: all the source words that align to the target one, represented in BIO[2] format.
- Source alignment context features: the combinations of the target word and one word before (left source context) or after (right source context) the source word aligned to it.

[2]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Statistics	WMT2014			WMT2013		
	train	**dev**	**test**	**train**	**dev**	**test**
#segments	1757	200	382	753	50	284
#words	40975	6436	9613	18435	1306	7827
%G (OK) : %B (BAD)	67 : 33	58 : 42	-	70 : 30	77 : 23	75 : 25

Table 1: Statistics of corpora used in LIG's system. We use the notion name+year to indicate the dataset. For instance, **train14** stands for the training set of WMT14

- Target alignment context features: the combinations of the source word and each word in the window ± 2 (two before, two after) of the target word.
- Backoff Behaviour: a score assigned to the word according to how many times the target Language Model has to back-off in order to assign a probability to the word sequence, as described in (Raybaud et al., 2011).
- Part-Of-Speech (POS) features (using TreeTagger[3] toolkit): The target word's POS; the source POS (POS of all source words aligned to it); bigram and trigram sequences between its POS and the POS of previous and following words.
- Binary lexical features that indicate whether the word is a: *stop word* (based on the stop word list for target language), *punctuation* symbol, *proper name* or *numerical*.
- Language Model (LM) features: the *"longest target n-gram length"* and *"longest source n-gram length"*(length of the longest sequence created by the current target (source aligned) word and its previous ones in the target (source) LM). For example, with the target word w_i: if the sequence $w_{i-2}w_{i-1}w_i$ appears in the target LM but the sequence $w_{i-3}w_{i-2}w_{i-1}w_i$ does not, the n-gram value for w_i will be 3.
- The *word's constituent label* and *its depth in the tree* (or the distance between it and the tree root) obtained from the constituent tree as an output of the Berkeley parser (Petrov and Klein, 2007) (trained over a Spanish treebank: AnCora[4]).
- Occurrence in Google Translate hypothesis: we check whether this target word appears in the sentence generated by Google Translate engine for the same source.
- Polysemy Count: the *number of senses* of each word given its POS can be a reliable indicator for judging if it is the translation of a particular source word. Here, we investigate the polysemy characteristic in both target word and its aligned source word. For source word (English), the number of senses can be counted by applying a Perl extension named Lingua::WordNet[5], which provides functions for manipulating the WordNet database. For target word (Spanish), we employ BabelNet[6] - a multilingual semantic network that works similarly to WordNet but covers more European languages, including Spanish.

3.2 WMT14 additional features

- POS's LM based features: we exploit the Spanish and English LMs of POS tag (provided as additional resources for this year's QE tasks) for calculating the maximum length of the sequences created by the current target token's POS and those of previous ones. The same score for POS of aligned source word(s) is also computed. Besides, the back-off score for word's POS tag is also taken into consideration. Actually, these feature types are listed in Section 3.1 for target word, and we proposed the similar ones for POS tags. In summary, three POS LM's new features are built, including: *"longest target n-gram length"*, *"longest source n-gram length"* and *back-off score* for POS tag.
- Word Occurrence in multiple translations: one novel point in this year's shared task is that the targets come from multiple MT

[3]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[4]http://clic.ub.edu/corpus/en/ancora
[5]http://search.cpan.org/dist/Lingua-Wordnet/Wordnet.pm
[6]http://babelnet.org

outputs (from systems or from humans) for the same source sentences. Obviously, one would have a "natural" intuition that: the occurrence of a word in all (or almost) systems implies a higher likelihood of being a correct translation. Relying on this observation, we add a new binary-value feature, telling whether the current token can be found in more than **N%** (in our experiments, we choose $N = 50$) out of all translations generated for the same source sentence. Here, in order to make the judgments more accurate, we propose several additional references besides those provided in the corpora, coming from: (1) Google Translate system, (2) The baseline SMT engine provided for WMT2013 English - Spanish QE task. These two MT outputs are added to the already available MT outputs of a given source sentence, before calculating the (above described) binary feature.

4 Baseline Experiments and Optimization Strategies

4.1 Machine Learning Method

Motivated by the idea of addressing Word Confidence Estimation (WCE) problem as a sequence labeling process, we employ the Conditional Random Fields (CRFs) for our model training, with WAPITI toolkit (Lavergne et al., 2010). Let $X = (x_1, x_2, \ldots, x_N)$ be the random variable over data sequence to be labeled, $Y = (y_1, y_2, \ldots, y_N)$ be the output sequence obtained after the labeling task. Basically, CRF computes the probability of the output sequence Y given the input sequence X by:

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} exp\left\{\sum_{k=1}^{K} \theta_k F_k(X, Y)\right\} \quad (1)$$

where $F_k(X, Y) = \sum_{t=1}^{T} f_k(y_{t-1}, y_t, x_t)$; $\{f_k\}(k = \overline{1, K})$ is a set of feature functions; $\{\theta_k\}(k = \overline{1, K})$ are the associated parameter values; and $Z_\theta(x)$ is the normalization function. In the training phase, we set the maximum number of iterations, the stop window size, and stop epsilon value at $200; 6$ and 0.00005 respectively.

System	Label	Pr(%)	Rc(%)	F(%)
BL(bin)	OK	66.67	81.92	73.51
	Bad	60.69	41.92	49.58
BL(L1)	OK	63.86	82.83	72.12
	Accuracy	22.14	14.89	17.80
	Fluency	50.40	27.98	35.98
BL(mult)	OK	63.32	87.56	73.49
	Fluency	14.44	10.10	11.88
	Mistranslation	9.95	5.69	7.24
	Terminology	3.62	3.89	3.75
	Unintelligible	52.97	16.56	25.23
	Agreement	5.93	11.76	7.88
	Untranslated	5.65	7.76	6.53
	Punctuation	56.97	25.82	35.53
BL+WMT 13(bin)	OK	68.62	82.69	75.01
	Bad	64.38	45.73	53.47

Table 2: Average Pr, Rc and F for labels of all-feature binary and multi-class systems, obtained on our WMT 2014 dev set (200 sentences). In **BL(multi)**, classes with zero value for Pr or Rc will not be reported

4.2 Experimental Classifiers

We experiment with the following classifiers:

- **BL(bin)**: all features (WMT14+WMT13) trained on **train14** only, using binary labels (*"OK"* and *"BAD"*)
- **BL(L1)**: all features trained on **train14** only, using level 1 labels (*"OK"*, *"Accuracy"*, and *"Fluency"*)
- **BL(mult)**: all features trained on **train14** only, using 16 labels
- **BL+WMT13(bin)**: all features trained on **train14 +** {**train+dev+test**}**13**, using binary labels.

System quality in Precision (Pr), Recall (Rc) and F score (F) are shown in Table 2. It can be observed that promising results are found in binary variant where both **BL(bin)** and **BL+WMT(bin)** are able to reach at least 50% F score in detecting errors (*BAD* class), meanwhile the performances in *"OK"* class go far beyond (73.51% and 75.01% respectively). Interestingly, the combination with WMT13 data boosts the baseline prediction capability in both labels: **BL+WMT13(bin)** outperforms **BL(bin)** in 1.10% (3.89%) for *OK* (*BAD*) label. Nevertheless, level 1 and multi-class systems maintain only good score for *"OK"* class. In addition, **BL(mult)** seems suffer seriously from its class imbalance, as well as the lack of training data for each, resulting in the inability of prediction for several among them (not all are reported in Table 2).

4.3 Decision threshold tuning for binary task

In binary systems **BL(bin)** and **BL+WMT13(bin)**, we run the classification task multiple times, corresponding to a decision threshold increase from 0.300 to 0.975 (step = 0.025). The values of Precision (Pr), Recall (Rc) and F-score (F) for *OK* and *BAD* label are tracked along this threshold variation, allowing us to select the optimal threshold that yields the highest $F_{avg} = \frac{F(OK)+F(BAD)}{2}$. Figure 1 shows that **BL(bin)** reaches the best performance at the threshold value of **0.95**, meanwhile the one for **BL+WMT13(bin)** is **0.75**. The latter threshold (0.75) has been used for the primary system submitted.

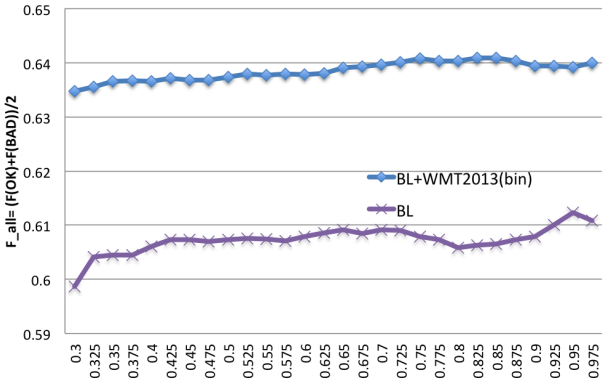

Figure 1: Decision threshold tuning on **BL(bin)** and **BL+WMT2013(bin)**

4.4 Feature Selection

In order to improve the preliminary scores of all-feature systems, we conduct a feature selection which is based on the hypothesis that some features may convey "noise" rather than "information" and might be the obstacles weakening the other ones. In order to prevent this drawback, we propose a method to filter the best features based on the "Sequential Backward Selection" algorithm[7]. We start from the full set of N features, and in each step sequentially remove the most useless one. To do that, all subsets of (N-1) features are considered and the subset that leads to the best performance gives us the weakest feature (not involved in the considered set). This procedure is also called "leave one out" in the literature. Obviously, the discarded feature is not considered in the following steps. We iterate the process until there is only one remaining feature in the set, and use the following score for comparing systems: $F_{avg}(all) = \frac{F_{avg}(OK)+F_{avg}(BAD)}{2}$, where $F_{avg}(OK)$ and $F_{avg}(BAD)$ are the averaged F scores for *OK* and *BAD* label, respectively, when threshold varies from 0.300 to 0.975. This strategy enables us to sort the features in descending order of importance, as displayed in Table 3. Figure 2 shows the evolution of the performance as more and more features are removed. The feature selection is done from the **BL+WMT2013(bin)** system.

We observe in Table 3 four valuable features which appear in top 10 in both WMT13 and WMT14 systems: *Source POS*, *Occur in Google Translate*, *Left source context* and *Right target context*. Among our proposed features, *"Occurrence in multiple systems"* is the most outstanding one with rank 3, *"longest target POS gram length"* plays an average role with rank 12, whereas *"longest source POS gram length"* is much less beneficial with the last position in the list. Figure 2 reveals that the optimal subset of features is the top 18 in Table 3, after discarding 6 weakest ones. This set will be used to train again the classifiers in all subtasks and compare to the baseline ones.

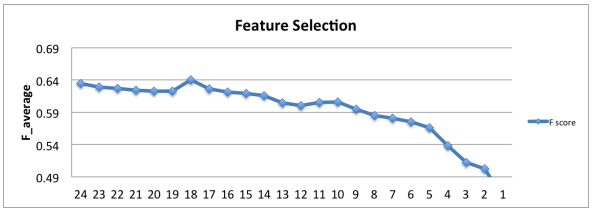

Figure 2: The evolution of the performance as more and more features are removed (from **BL+WMT2013(bin)** system)

5 Submissions

After finishing the optimization process and comparing systems, we select two most out-standing ones (of each subtask) for the submission of this year's shared task. They are the following:

- Binary variant: **BL+WMT13(bin)** and **FS(bin)** (feature selection from the same corresponding system)
- Level 1 variant: **BL(L1)** and **FS(L1)** (feature selection from the same corresponding system)

[7] http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf

Rank	WMT2014	WMT2013
1	Target POS	**Source POS**
2	Longest target gram length	**Occur in Google Translate**
3	Occurrence in multiple systems	Nodes
4	Target word	Target POS
5	**Occur in Google Translate**	WPP any
6	**Source POS**	**Left source context**
7	Numeric	**Right target context**
8	Polysemy count (target)	Numeric
9	**Left source context**	Polysemy count(target)
10	**Right Target context**	Punctuation
11	Constituent label	**Stop word**
12	Longest target POS gram length	Right source context
13	Punctuation	**Target word**
14	Stop word	Distance to root
15	Number of occurrences	Backoff behaviour
16	Left target context	Constituent label
17	Backoff behaviour	Proper name
18	Polysemy count (source)	Number of occurrences
19	Source Word	Min
20	Proper Name	Max
21	Distance to root	Left target context
22	Longest source gram length	Polysemy count (source)
23	Right source context	Longest target gram length
24	Longest source POS gram length	Longest source gram length
25		Source Word

Table 3: The rank of each feature (in term of usefulness) in **WMT2014** and **WMT2013** systems. The bold ones perform well in both cases. Note that feature sets are not exactly the same for 2013 and 2014 (see explanations in section 3).

- Multi-class variant: **BL(mult)** and **FS(mult)** (feature selection from the same corresponding system)

The official results can be seen in Table 4. This year, in order to appreciate the translation error detection capability of WCE systems, the **official** evaluation metric used for systems ranking is the **average F score** for all but the *"OK"* class. For the non-binary variant, this average is weighted by the frequency of the class in the test data. Nevertheless, we find the F scores for *"OK"* class are also informative, since they reflect how good our systems are in identifying correct translations. Therefore, both scores are reported in Table 4.

System	F("OK") (%)	Average F(%)
FS(bin) (primary)	74.0961	0.444735
FS(L1)	73.9856	0.317814
FS(mult)	76.6645	0.204953
BL+WMT2013(bin)	74.6503	0.441074
BL(L1)	74.0045	0.317894
BL(mult)	76.6645	0.204953

Table 4: The F scores for "OK" class and the average F scores for the remaining classes (official WMT14 metric) , obtained on test set.

6 Conclusion and perspectives

We presented our preparation for this year's shared task on QE at word level, for the English - Spanish language pair. The lack of some information on MT system internals was a challenge. We made efforts to maintain most of well-performing 2013 features, especially the source side ones, and propose some novel features based on this year's corpus specificities, as well as combine them with those of last year. Generally, our results are not able to beat those in WMT13 for the same language pair, yet still promising under these constraints. As future work, we are thinking of using more efficiently the existing references (coming from provided translations and other reliable systems) to obtain stronger indicators, as

well as examine other ML methods besides CRF.

References

Ergun Bicici. Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-2242`.

Aaron Li-Feng Han, Yi Lu, Derek F. Wong, Lidia S. Chao, Liangye He, and Junwen Xing. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 365–372, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-2245`.

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, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic, June 2007.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting et labeling sequence data. In *Proceedings of ICML-01*, pages 282–289, 2001.

Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513, 2010.

Ngoc Quang Luong, Laurent Besacier, and Benjamin Lecouteux. Word confidence estimation and its integration in sentence quality estimation for machine translation. In *Proceedings of the fifth international conference on knowledge and systems engineering (KSE)*, Hanoi, Vietnam, October 2013.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19–51, 2003.

Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, NY, April 2007.

S. Raybaud, D. Langlois, and K. Smaï li. "this sentence is wrong." detecting errors in machine - translated sentences. In *Machine Translation*, pages 1–34, 2011.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Terp system description. In *MetricsMATR workshop at AMTA*, 2008.