Word Confidence Estimation for SMT N-best List Re-ranking

Ngoc-Quang Luong, Laurent Besacier, Benjamin Lecouteux

Word Confidence Estimation for SMT *N*-best List Re-ranking

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 proposes to use Word Confidence Estimation (WCE) information to improve MT outputs via *N*-best list re-ranking. From the confidence label assigned for each word in the MT hypothesis, we add six scores to the baseline log-linear model in order to re-rank the *N*-best list. Firstly, the correlation between the WCE-based sentence-level scores and the conventional evaluation scores (BLEU, TER, TERp-A) is investigated. Then, the *N*-best list re-ranking is evaluated over different WCE system performance levels: from our real and efficient WCE system (ranked 1st during last WMT 2013 *Quality Estimation* Task) to an oracle WCE (which simulates an interactive scenario where a user simply validates words of a MT hypothesis and the new output will be automatically re-generated). The results suggest that our real WCE system slightly (but significantly) improves the baseline while the oracle one extremely boosts it; and better WCE leads to better MT quality.

1 Introduction

A number of methods to improve MT hypotheses after decoding have been proposed in the past, such as: post-editing, re-ranking or re-decoding. Post-editing (Parton et al., 2012) is a human-inspired task where the machine post edits translations in a second automatic pass. In re-ranking (Zhang et al., 2006; Duh and Kirchhoff, 2008; Bach et al., 2011), more features are used along with the multiple model scores for re-determining the 1-best among *N*-best list. Meanwhile, re-decoding process (Venugopal et al., 2007) intervenes directly into the decoder's search graph (e.g. adds more reward or penalty scores), driving it to another better path.

This work aims at re-ranking the *N*-best list to improve MT quality. Generally, during the translation task, the decoder traverses through paths in its search space, computes the objective function values for them and outputs the one with highest score as the best hypothesis. Besides, those with lower scores can also be generated in a so-called *N*-best list. The decoder's function consists of parameters from different models, such as translation, distortion, word penalties, reordering, language models, etc. In the *N*-best list, although the current 1-best beats the remains in terms of model score, it might not be exactly the closest to the human reference. Therefore, adding more decoder independent features would be expected to raise up a better candidate. In this work, we build six additional features based on the labels predicted by our Word Confidence Estimation (WCE) system, then integrate them with the existing decoder scores for re-ranking hypotheses in the *N*-best list. More precisely, ***in the second pass***, our re-ranker aggregates over decoder and WCE-based weighted scores and utilizes the obtained sum to sort out the best candidate. The novelty of this paper lies on the following contributions: the correlation between WCE-based sentence-level scores and conventional evaluation scores (BLEU, TER, TERp-A) is first investigated. Then, we conduct the *N*-best list re-ranking over different WCE system performance levels: starting by a real WCE, passing through several gradually improved (simulated) systems and finally the "oracle" one. From these in-depth experiments, the role of WCE in improving MT quality via re-ranking *N*-best list is confirmed and reinforced.

The remaining parts of this article are organized as follows: in section 2 we summarize some outstanding approaches in *N*-best list re-ranking as well as in WCE. Section 3 describes our WCE system construction, followed by proposed features.

The experiments along with results and in-depth analysis of WCE scores' contribution (as WCE system gets better) are presented in Section 4 and Section 5. The last section concludes the paper and points out some ongoing work.

2 Related Work

2.1 *N*-best List Re-ranking

Walking through various related work concerning this issue, we observe some prominent ideas. The first attempt focuses on proposing additional Language Models. Kirchhoff and Yang (2005) train one word-based 4-gram model (with modified Kneser-Ney smoothing) and one factored trigram one, then combine them with seven decoder scores for re-ranking *N*-best lists of several SMT systems. Their proposed LMs increase the translation quality of the baselines (measured by BLEU score) from 21.6 to 22.0 (Finnish - English), or from 30.5 to 31.0 (Spanish - English). Meanwhile, Zhang et al. (2006) experiment a distributed LM where each server, among the total of 150, hosts a portion of the data and responses its client, allowing them to exploit an extremely large corpus (2.7 billion word English Gigaword) for estimating N-gram probability. The quality of their Chinese - English hypotheses after the re-scoring process by using this LM is improved 4.8% (from BLEU 31.44 to 32.64, oracle score = 37.48).

In one other direction, several authors propose to replace the current linear scoring function used by the decoder by more efficient functions. Sokolov et al. (2012) learn their non-linear scoring function in a learning-to-rank paradigm, applying Boosting algorithm. Their gains on the WMT'{10, 11, 12} are shown modest yet consistent and higher than those based on linear scoring functions. Duh and Kirchhoff (2008) use Minimum Error Rate Training (MERT) (Och, 2003) as a weak learner and build their own solution, BoostedMERT, a highly-expressive re-ranker created by voting among multiple MERT ones. Their proposed model dramatically beats the decoder's log-linear model (43.7 vs. 42.0 BLEU) in IWSLT 2007 Arabic - English task. Applying solely *goodness* (the sentence confidence) scores, Bach et al. (2011) obtain very consistent TER reductions (0.7 and 0.6 on the dev and test set) after a 5-list re-ranking for their Arabic - English SMT hypotheses. This latter work is the one that is the most related to our paper. However, the major differences are: (1) our proposed sentence scores *are computed based on word confidence labels*; and (2) we perform an in-depth study of the use of WCE for N-best reranking and assess its usefulness in a simulated interactive scenario.

2.2 Word Confidence Estimation

Confidence Estimation (CE) is the task of identifying the correct parts and detecting the translation errors in MT output. If the error is predicted for each word, this becomes WCE. The interesting uses of WCE include: pointing out the words that need to be corrected by the post-editor, telling readers about the reliability of a specific portion, and selecting the best segments among options from multiple translation systems for combination.

Dealing with this problem, various approaches have been proposed: Blatz et al. (2003) combine several features using neural network and naive Bayes learning algorithms. One of the most effective feature combinations is the Word Posterior Probability (WPP) as suggested by Ueffing et al. (2003) associated with IBM-model based features (Blatz et al., 2004). Ueffing and Ney (2005) propose an approach for phrase-based translation models: a phrase is a sequence of contiguous words and is extracted from the word-aligned bilingual training corpus. The confidence value of each word is then computed by summing over all phrase pairs in which the target part contains this word. Xiong et al. (2010) integrate target word's Part-Of-Speech (POS) and train them by Maximum Entropy Model, allowing significative gains in comparison to WPP features. The novel features from source side, alignment context, and dependency structure (Bach et al., 2011) help to augment marginally in F-score as well as the Pearson correlation with human judgment. Other approaches are based on external features (Soricut and Echihabi, 2010; Felice and Specia, 2012) allowing to cope with various MT systems (e.g. statistical, rule based etc.). Among the numerous WCE applications, we consider its contribution in a specific step of SMT pipeline: *N*-best list re-ranking. Our WCE system and the proposed re-ranking features are presented in the next section.

3 Our Approach

Our approach can be expressed in three steps: investigate the potential of using word-level score in *N*-best list re-ranking, build the WCE system and

extract additional features to integrate with the existing log-linear model.

3.1 Investigating the correlation between "word quality" scores and other metrics

Firstly, we investigate the correlation between sentence-level scores (obtained from WCE labels) and conventional evaluation scores (BLEU (Papineni et al., 2002), TER and TERp-A (Snover et al., 2008)). For each sentence, a word quality score (WQS) is calculated by:

$$WQS = \frac{\#''G''(good)\ words}{\#words} \quad (1)$$

In other words, we are trying to answer the following question: can the high percentage of *"G"* (good) words (predicted by WCE system) in a MT output ensure its possibility of having a better BLEU and low TER (TERp-A) value ? This investigation is a strong prerequisite for further experiments in order to check that WCE scores do not bring additional "noise" to the re-ranking process. In this experiment, we compute WQS over our entire French - English data set (total of 10,881 1-best translations) for which WCE **oracle labels** are available (see Section 3.2 to see how they were obtained). The results are plotted in Figure 1, where the *y* axis shows the *"G"* (good) word percentage, and the *x* axis shows BLEU (1a), TER (1b) or TERp-A (1c) scores. It can be seen from Figure 1 that the major parts of points (the densest areas) in all three cases conform the common tendency: In Figure 1a, the higher "G" percentage, the higher BLEU is; on the contrary, in Figure 1b (Figure 1c), the higher "G" percentage, the lower TER (TERp-A) is. We notice some outliers, i.e. sentences with most or almost words labeled "good", yet still have low BLEU or high TER (TERp-A) scores. This phenomenon is to be expected when many (unknown) source words are not translated or when the (unique) reference is simply too far from the hypothesis. Nevertheless, the information extracted from oracle WCE labels seems useful to build an efficient re-ranker.

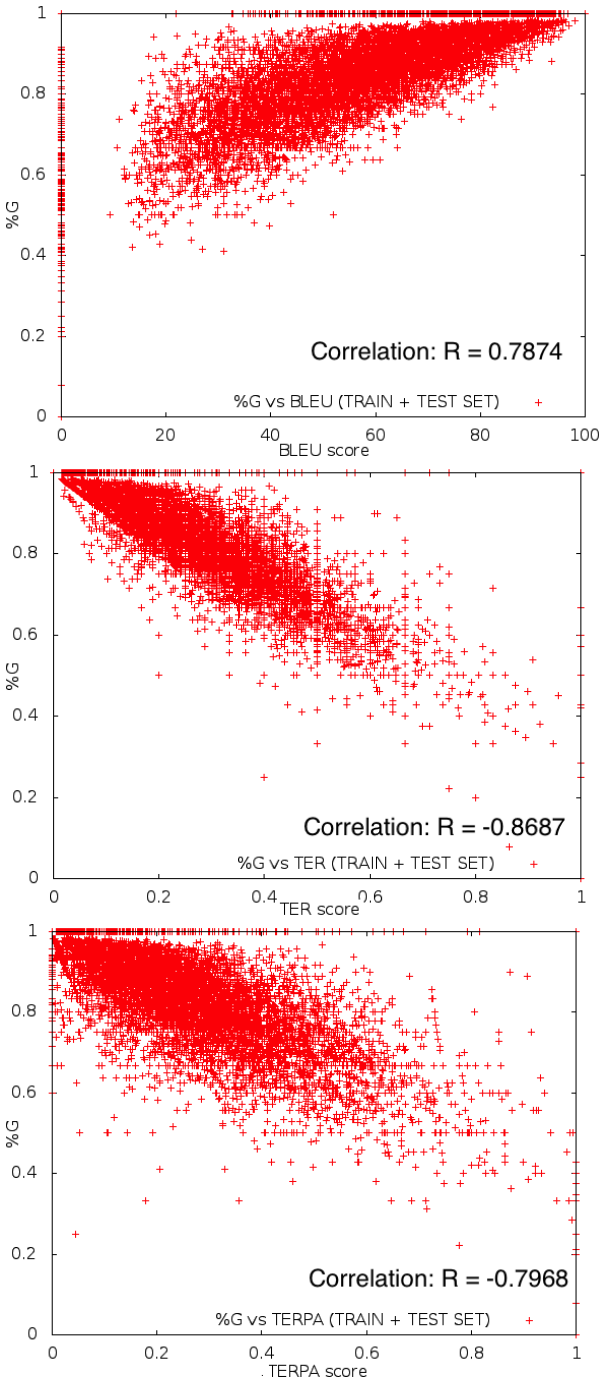

Figure 1: The correlation between WQS in a sentence and its overall quality measured by : (a) BLEU, (b) TER and (c) TERp-A metrics

3.2 WCE System Preparation

Essentially, a WCE system construction consists of two pivotal elements: the features (the SMT system dependent or independent information extracted for each word to represent its characteristics) and the machine learning method (to train the prediction model). Motivated by the idea of addressing 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). Basically, CRF computes the probability of the output sequence $Y = (y_1, y_2, ..., y_N)$ given the input sequence $X = (x_1, x_2, ..., x_N)$ by:

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} exp\left\{\sum_{k=1}^{K} \theta_k F_k(X,Y)\right\} \quad (2)$$

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 terms of features, a number of knowledge sources are employed for extracting them, resulting in the major types listed below. We briefly summarize them in this work, further details about total of 25 features can be referred in (Luong et al., 2013a).

- Target Side: target word; bigram (trigram) backward sequences; number of occurrences
- Source Side: source word(s) aligned to the target word
- Alignment Context: the combinations of the target (source) word and all aligned source (target) words in the window ± 2
- Word posterior probability
- Pseudo-reference (Google Translate): whether the current word appears in the pseudo reference or not[1]?
- Graph topology: number of alternative paths in the confusion set, maximum and minimum values of posterior probability distribution
- Language model (LM) based: length of the longest sequence of the current word and its previous ones in the target (resp. 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.
- Lexical Features: word's Part-Of-Speech (POS); sequence of POS of all its aligned source words; POS bigram (trigram) backward sequences; punctuation; proper name; numerical
- Syntactic Features: Null link; constituent label; depth in the constituent tree
- Semantic Features: number of word senses in WordNet.

Interestingly, this feature set was also used in our English - Spanish WCE system which got the first rank in WMT 2013 Quality Estimation Shared Task (Luong et al., 2013b).

For building the WCE training and test sets, we use a dataset of 10,881 French sentences (Potet et al., 2012) , and apply a baseline SMT system to generate hypotheses (1000-best list). Our baseline SMT system (presented for WMT 2010 evaluation) keeps the Moses's default setting (Koehn et al., 2007): log-linear model with 14 weighted feature functions. The translation model is trained on the Europarl and News parallel corpora of WMT10[2] evaluation campaign (1,638,440 sentences). The target language model is trained by the SRI language modeling toolkit (Stolcke, 2002) on the news monolingual corpus (48,653,884 sentences).

Translators were then invited to correct MT outputs, giving us the same amount of post editions (Potet et al., 2012). The set of triples (source, hypothesis, post edition) is then divided into the training set (10000 first triples) and test set (881 remaining). To train the WCE model, we extract all above features for words of the **1-best hypotheses** of the training set. For the test set, the features are built **for all 1000 best translations** of each source sentence. Another essential element is the word's confidence labels (or so-called WCE oracle labels) used to train the prediction model as well as to judge the WCE results. They are set by using TERp-A toolkit (Snover et al., 2008) in one of the following classes: "I' (insertions), "S" (substitutions), "T" (stem matches), "Y" (synonym matches), "P" (phrasal substitutions), "E" (exact matches) and then simplified into binary class: *"G"* (good word) or *"B"* (bad word) (Luong et al., 2013a).

Once having the prediction model built with all features, we apply it on the test set (881 x 1000 best = 881000 sentences) and get needed WCE labels. Figure 2 shows an example about the classification results for one sentence. Comparing with the reference labels, we can point out easily the correct classifications for *"G"* words (e.g. in case of *operation*, *added*) and for *"B"* words (e.g. *is*, *have*), as well as classification errors (e.g. *a*, *combat*). According to the Precision (Pr), Recall (Rc) and F-score (F) shown in Table 1, our WCE system reaches very promising performance in predicting *"G"* label, and acceptable for *"B"* label. These labels will be used to calculate our proposed

[1]This is our first-time experimented feature and does not appear in (Luong et al., 2013a)

[2]http://www.statmt.org/wmt10/

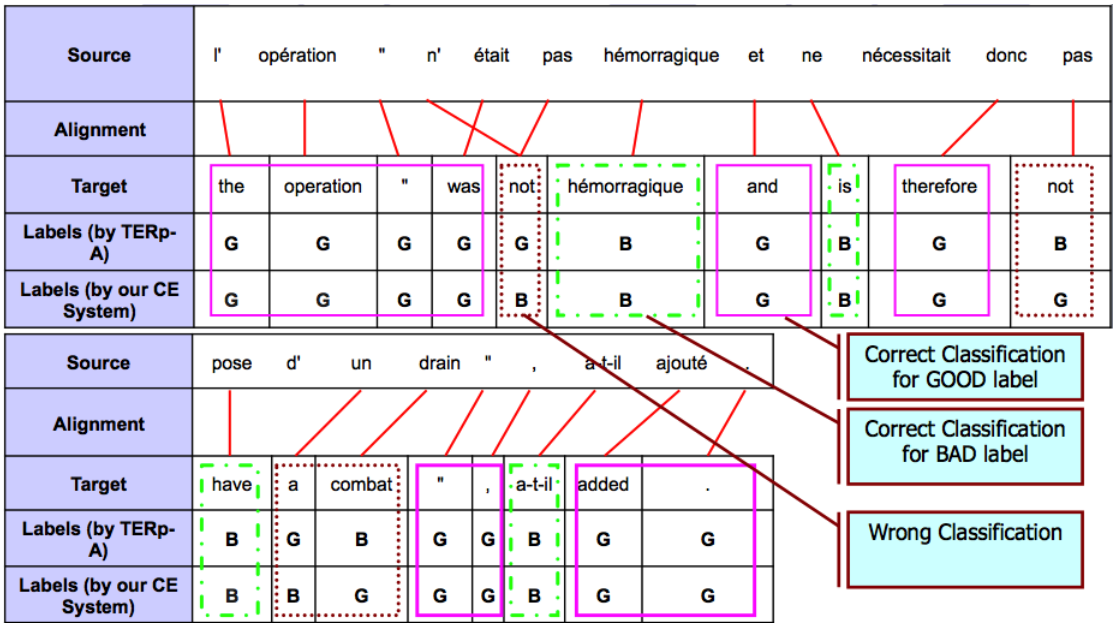

Figure 2: Example of our WCE classification results for one MT hypothesis

features (section 3.3).

Label	Pr(%)	Rc(%)	F(%)
Good (G)	84.36	91.22	**87.65**
Bad (B)	51.34	35.95	**42.29**

Table 1: Pr, Rc and F for "G" and "B" labels of our WCE system

3.3 Proposed Features

Since the scores resulted from the WCE system are for words, we have to synthesize them in sentence level scores for integrating with the 14 decoder scores. Six proposed scores involve:

- The ratio of number of good words to total number of words. (1 score)
- The ratio of number of good nouns (verbs) to total number of nouns (verbs)[3]. (2 scores)
- The ratio of number of n consecutive good word sequences to the total number of consecutive word sequences ; n=2, n=3 and n=4. (3 scores)

For instance, in case of the hypothesis in Figure 2: among the total of 18 words, we have 12 labeled as *"G"*; and 7 out of 17 word pairs (bigram) are labeled as *"GG"*, etc. Hence, some of the above scores can be written as:

$$\begin{aligned} \frac{\#good\ words}{\#words} &= \frac{12}{18} = 0.667 \\ \frac{\#good\ bigrams}{\#bigrams} &= \frac{7}{17} = 0.4118 \\ \frac{\#good\ trigrams}{\#trigrams} &= \frac{3}{16} = 0.1875 \end{aligned} \quad (3)$$

With the features simply derived from WCE labels and not from CRF model scores (i.e. the probability $p(G)$, $p(B)$) , we expect to spread out the evaluation up to the "oracle" setting, where the users validate a word as "G" or "B" without providing any confidence score.

4 Experiments

4.1 Experimental Settings

As described in Section 3.2, our SMT system generates 1000-best list for each source sentence, and among them, the best hypothesis was determined by using the objective function based on 14 decoder scores, including: 7 reordering scores, 1 language model score, 5 translation model scores and 1 word penalty score. Initially, all six additional WCE-based scores are weighted as 1.0. Then, two optimization methods: MERT and Margin Infused Relaxed Algorithm (MIRA) (Watanabe et al., 2007) are applied to optimize the weights of all 20 scores of the re-ranker. In both methods, we carry out a **2-fold cross validation** on the *N*-best

[3]We decide not to experiment with adjectives, adverbs and conjunctions since their number can be 0 in many cases.

Systems	MERT			MIRA		
	BLEU	TER	TERp-A	BLEU	TER	TERp-A
BL	52.31	0.2905	0.3058	50.69	0.3087	0.3036
BL+OR	**58.10**	**0.2551**	**0.2544**	**55.41**	**0.2778**	**0.2682**
BL+WCE	52.77	0.2891	0.3025	51.01	0.3055	0.3012
WCE + 25%	53.45	0.2866	0.2903	51.33	0.3010	0.2987
WCE + 50%	55.77	0.2730	0.2745	53.63	0.2933	0.2903
WCE + 75%	56.40	0.2687	0.2669	54.35	0.2848	0.2822
Oracle computed from N-best translations	**BLEU=60.48**					

Table 2: Translation quality of the baseline system (only decoder scores) and that with additional scores from real "WCE" or "oracle" WCE system

System	MERT		
	Better	Equivalent	Worse
BL+WCE	159	601	121
BL+OR	**517**	261	153
WCE+25%	253	436	192
WCE+50%	320	449	112
WCE+75%	461	243	177

Table 3: Quality comparison (measured by TER) between the baseline and two integrated systems in details (How many sentences are improved, kept equivalent or degraded, out of 881 test sentences?

test set. In other words, we split our *N*-best test set into two equivalent subsets: S1 and S2. Playing the role of a development set, S1 will be used to optimize the 20 weights for re-ranking S2 (and vice versa). Finally two result subsets (new 1-best after re-ranking process) are merged for evaluation. To better acknowledge the impact of the proposed scores, we calculate them not only using our real WCE system, but also using an oracle WCE (further called "WCE scores" and "oracle scores", respectively). To summarize, we experiment with the three following systems:

- **BL**: Baseline SMT system with 14 above decoder scores
- **BL+WCE**: Baseline + 6 real WCE scores
- **BL+OR**: Baseline + 6 oracle WCE scores (simulating an interactive scenario).

4.2 Results and Analysis

The translation quality of **BL**, **BL+WCE** and **BL+OR**, optimized by MERT and MIRA method are reported in Table 2. Meanwhile, Table 3 depicts in details the number of sentences in the two integrated systems which outperform, remain equivalent or degrade the baseline hypothesis (when match against the references, measured by TER). It can be observed from Table 2 that the integration of **oracle scores** significantly boosts the MT output quality, measured by all three metrics and optimized by both methods employed. We gained 5.79 and 4.72 points in BLEU score, by MERT and MIRA (respectively). With TER, **BL+OR** helps to gain 0.03 point in both two methods. Meanwhile, in case of TERp-A, the improvement is 0.05 point for MERT and 0.03 point for MIRA. It is worthy to mention that the possibility of obtaining such oracle labels is definitely doable through a human-interaction scenario (which could be built from a tool like PET (Post-Editing Tool) (Aziz et al., 2012) for instance). In such an environment, ***once having the hypothesis produced by the first pass (translation task)***, the human editor could simply click on words considered as bad (B), the other words being implicitly considered as correct (G).

Breaking down the analysis into sentence level, as described on Table 3, **BL+OR** (MERT) yields nearly 59% (517 over 881) better outputs than the baseline and only 17% of worse ones. Furthermore, Table 2 shows that in case of our test set, optimizing by MERT is pretty more beneficial than MIRA (we do not have a clear explanation of this yet).

For more insightful understanding about WCE scores' acuteness, we make a comparison with

the most possible optimal BLEU score that could be obtained from the *N*-best list. Applying the sentence-level BLEU+1 (Nakov et al., 2012) metric over candidates in the list, we are able to select the one with highest score and aggregate all of them in an oracle-best translation; the resulting performance obtained is **60.48**. This score accounts for a fact that the simulated interactive scenario (**BL+OR**) lacks only 2.38 points (in case of MERT) to be optimal and clearly overpass the baseline (8.17 points below the best score).

The contribution of a real WCE system seems more modest: **BL+WCE** marginally increases BLEU scores of **BL** (0.46 gain in case of optimizing by MERT and 0.32 by MIRA). For both TER and TERp-A metric, the progressions are also negligible. To verify the significance of this result, we estimate the *p*-value between BLEU of **BL+WCE** system and BLEU of baseline **BL** relying on Approximate Randomization (AR) method (Clark et al., 2011) which indicates if the improvement yielded by the optimized system is likely to be generated again by some random processes (randomized optimizers). After various optimizer runs, we selected randomly *5* optimizer outputs, perform the AR test and obtain a *p*-value of **0.01**. This result reveals that the improvement yielded by **BL+WCE** is significative although small, originated from the contribution of WCE score, not by any optimizer variance. This modest but positive change in BLEU score using WCE features, encourages us to investigate and analyze further about WCE scores' impact, supposing WCE performance is getting better. More in-depth analysis is presented in the next section.

5 Further Understanding of WCE scores role in *N*-best Re-ranking via Improvement Simulation

We think it would be very interesting and useful to answer the following question: do WCE scores really effectively help to increase MT output quality when the WCE system is getting better and better? To do this, our proposition is as follows: firstly, by using the oracle labels, we filter out all wrongly classified words in the test set and push them into a temporary set, called **T**. Then, we correct randomly a percentage (25%, 50%, or 75%) of labels in **T**. Finally, the altered **T** will be integrated back with the correctly predicted part (by the WCE system) in order to form a new "simulated" result set. This strategy results in three "virtual" WCE systems called **"WCE+N%"** (N=25, 50 or 75), which use 14 decoder scores and 6 "simulated" WCE scores. Table 4 shows the performance of these systems in term of F score (%). From each of the above systems, the whole experimental setting is identical to what we did with the original WCE and oracle systems: six scores are built and combined with existing 14 system scores for each hypothesis in the *N*-best list. After that, MERT and MIRA methods are invoked to optimize their weights, and finally the reordering is performed thanks to these scores and appropriate optimal weights. The translation quality measured by BLEU, TER and TERp-A after re-ranking using **"WCE+N%"** (N=25,50,75) can be seen also in Table 2. The number of translations which outperform, keep intact and decline in comparison to the baseline are shown in Table 3 for MERT optimization.

System	F("G")	F("B")	Overall F
WCE+25%	89.87	58.84	63.51
WCE+50%	93.21	73.09	76.11
WCE+75%	96.58	86.87	88.33
Oracle labels	100	100	100

Table 4: The performances (Fscore) of simulated WCE systems

We note that all obtained scores fit our guess and expectation: the better performance WCE system reaches, the clearer its role in improving MT output quality. Diminishing 25% of the wrongly predicted words leads to a gain 0.68 point (by MERT) and 0.32 (by MIRA) in BLEU score. More significant increases of BLEU 3.00 and BLEU 3.63 (MERT) can be achieved when prediction errors are cut off up to 50% and 75%. Figure 3 presents an overview of the results obtained and helps us to predict the MT improvements expected if the WCE system improves in the future. Table 5 shows several examples where WCE scores drive SMT system to better reference-correlated hypothesis. In the first example, the baseline generates the hypothesis in which the source phrase *"pour sa part"* remains untranslated. On the contrary, **WCE+50%** overcomes this drawback by resulting in a correct translation phrase: *"for his part"*. The latter translation needs only one edit operation (shift for *"Bettencourt-Meyers"*) to become its reference. In example 2, **BL+OR** selects the

Example 1 (from WCE+50%)	
Source	Pour sa part , l' avocat de Françoise Bettencourt-Meyers , Olivier Metzner , s' est félicité de la décision du tribunal .
Hypothesis (Baseline SMT)	The lawyer of *Bettencourt-Meyers* Françoise , Olivier Metzner , welcomed the court 's decision .
Hypothesis (SMT+WCE scores)	*For his part ,* the lawyer of *Bettencourt-Meyers* Françoise , Olivier Metzner , welcomed the court 's decision .
Post-edition	*For his part ,* the lawyer of Françoise Bettencourt-Meyers , Olivier Metzner , welcomed the court 's decision .
Example 2 (from BL+OR)	
Source	Pour l' otre , l' accord risque " de creuser la tombe d' un très grand nombre de pme du secteur dans les 12 prochains mois " .
Hypothesis (Baseline MT)	For the otre the agreement is likely to *deepen the grave* of a very large number of *smes in the sector* in the next 12 months " .
Hypothesis (SMT+WCE scores)	For the otre agreement , the risk " *digging the grave* of a very large number of *medium-sized businesses* in the next 12 months " .
Post-edition	For the otre , the agreement risks " *digging the grave* of a very large number of *small- and medium-sized businesses* in the next 12 months " .

Table 5: Examples of MT hypothesis before and after reranking using the additional scores from WCE+50% (Example 1) and BL+OR (Example 2) system

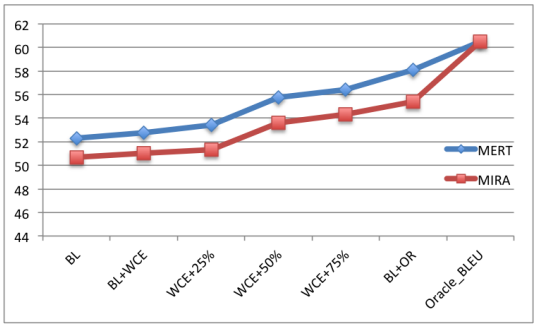

Figure 3: Comparison of the performance of various systems: the integrations of WCE features, which the quality increases gradually, lead to the linear improvement of translation outputs.

better hypothesis, in which the phrases *"creuser la tombe"* and *''pme du secteur"* are translated into *"digging the grave"* and *"medium-sized businesses"*, respectively, better than those of the baseline (*"deepen the grave"* and *"smes in the sector"*).

6 Conclusions And Perspectives

So far, the word confidence scores have been exploited in several applications, e.g. post-editing, sentence quality assessment or multiple MT-system combination, yet very few studies (except Bach et al. (2011)) propose to investigate them for boosting MT quality. Thus, this paper proposed several features extracted from a WCE system and combined them with existing decoder scores for re-ranking *N*-best lists. Our WCE model is built using CRFs, on a variety of types of features for the French - English SMT task. Due to its limitations in predicting translation errors ("B" label), WCE scores ensure only a modest improvement in translation quality over the baseline SMT. Nevertheless, further experiments about the simulation of WCE performance suggest that such types of score contribute dramatically if they are built from an accurate WCE system. They also show that with the help of an "ideal" WCE, the MT system reaches quite close to its most optimal possible quality. These scores are totally independent from the decoder, they can be seen as a way to introduce lexical, syntactic and semantic information (used for WCE) in a SMT pipeline.

As future work, we plan to focus on augmenting our WCE performance using more linguistic features as well as advanced techniques (feature selection, Boosting method...). In the same time, we would like to integrate the WCE scores in the decoder's search graph to redirect the decoding process (preliminary experiments, not reported here yet, have shown that this is a very promising avenue of research).

References

Wilker Aziz, Sheila C. M. de Sousa, and Lucia Specia. Pet: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 23-25 2012.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 211–219, Portland, Oregon, June 19-24 2011.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. Technical report, JHU/CLSP Summer Workshop, 2003.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Proceedings of COLING 2004*, pages 315–321, Geneva, April 2004.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the Association for Computational Lingustics*, 2011.

Kevin Duh and Katrin Kirchhoff. Beyond log-linear models: Boosted minimum error rate training for n-best re-ranking. In *Proc. of ACL, Short Papers*, 2008.

Mariano Felice and Lucia Specia. Linguistic features for quality estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 96–103, Montreal, Canada, June 7-8 2012.

Katrin Kirchhoff and Mei Yang. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, Ann Arbor, Michigan, June 2005.

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.

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 2013)*, Hanoi, Vietnam, October 17-19 2013a.

Ngoc Quang Luong, Benjamin Lecouteux, and Laurent Besacier. LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 396–391, Sofia, Bulgaria, August 2013b. Association for Computational Linguistics.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. Optimizing for sentence-level bleu+1 yields short translations. In *Proceedings of COLING 2012*, pages 1979–1994, Mumbai, India, December 8 -15 2012.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, July 2003.

Kishore Papineni, Salim Roukos, Todd Ard, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

Kristen Parton, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adrià de Gispert. Can automatic post-editing make mt more meaningful? In *Proceedings of the 16th EAMT*, pages 111–118, Trento, Italy, 28-30 May 2012.

M Potet, R Emmanuelle E, L Besacier, and H Blanchon. Collection of a large database of french-english smt output corrections. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May 2012.

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

Artem Sokolov, Guillaume Wisniewski, and Francois Yvon. Non-linear n-best list reranking with few features. In *Proceedings of AMTA*, 2012.

Radu Soricut and Abdessamad Echihabi. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th ACL (Association for Computational Linguistics)*, pages 612–621, Uppsala, Sweden, July 2010.

Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, USA, 2002.

Nicola Ueffing and Hermann Ney. Word-level confidence estimation for machine translation using phrased-based translation models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 763–770, Vancouver, 2005.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. Confidence measures for statistical machine translation. In *Proceedings of the MT Summit IX*, pages 394–401, New Orleans, LA, September 2003.

Ashish Venugopal, Andreas Zollmann, and Stephan Vogel. An efficient two-pass approach to synchronous-cfg driven statistical mt. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, April 2007.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 64–773,, Prague, Czech Republic, June 2007.

Deyi Xiong, Min Zhang, and Haizhou Li. Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th Association for Computational Linguistics*, pages 604–611, Uppsala, Sweden, July 2010.

Ying Zhang, Almut Silja Hildebrand, and Stephan Vogel. Distributed language modeling for n-best list re-ranking. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 216–223, Sydney, July 2006.