XXX*: an Open Source Toolkit for Word Confidence Estimation in Machine Translation

First Last
Address
{firstname.lastname}@address.com

Abstract

This paper presents an open-source toolkit for predicting the quality of words in a machine translation (MT) output whose novel contributions are (a) support for various language-pairs, (b) handle a number of features of different types (system-based, lexical, syntactic and semantic). In addition, the toolkit also integrates a wide variety of natural language processing or machine learning tools in order to preprocess data, extract features and estimate confidence at word-level. Features for WCE can be easily added / removed using a configuration file. We validate the toolkit by experimenting in the Quality Estimation (QE) evaluation framework of WMT with two bilingual corpora: French-English and English-Spanish. The toolkit is made available to the research community with ready-made scripts to launch full experiments on these language pairs, while achieving state-of-the-art performances. We also believe that the feature engineering part of our toolkit could be used for other NLP tasks.

1 Introduction

Statistical Machine Translation (SMT) has proven its efficiency during the last decade. For document translation, the following process is now broadly used: the SMT system produces raw translations then trained professional translators post-edit (correct) translation errors (PE). We believe that this SMT+PE pipeline can be improved using automatic confidence estimation (CE) where the system gives some clues about the quality of the translation output. For instance, post-editors require to have information about the possible quality of the translation (Should they just post-edit the translation or rewrite the whole output? What are the main words/phrases they need to focus on?).

Building a method, that could point out both correct and incorrect parts in MT output, is a key component to solve above problems. If we limit the concept 'parts' to 'words', that issue is called Word-level Confidence Estimation (WCE).

Past years have seen the emergence of shared tasks to estimate the translation quality (like WMT CE shared task[1]). In 2015, the organizers called for methods to predict the quality of MT output at run-time on three levels: sentence-level (Task 1), word-level (Task 2) and (new) document-level (Task 3). This paper more precisely deals with the second task (WCE) but we believe it might be of interest for any researcher working in quality assesment for MT.

Contributions Our experience in participating to *task 2* (WCE) in 2013 and 2014 lead us to the following observation: while feature processing is very important to achieve good performance, it requires to call a set of heterogeneous NLP tools (for lexical, syntactic, semantic analyses). Thus, we propose to unify the feature processing, together with the call of machine learning algorithms, in order to facilitate the design of confidence estimation systems. The open-source toolkit proposed (written in *Python* and made available on *github*) integrates some standard as well as in-house features that have proven useful for WCE (based on our experience in WMT 2013 and 2014).

To our knowledge, this is the first toolkit dedicated to word confidence estimation. In addition to describing the toolkit in the details, this paper shows how it can be easily adapted to new language pairs by just modifying a configuration file. We also report experiments on french-english and english-spanish tasks and show that the toolkit obtains similar (and state-of-the-art) performances compared to a previous system presented at WMT shared task in 2014.

We also believe that the integrated feature processing of our toolkit could be used for other cross-lingual NLP tasks.

2 WCE formalisation and related work

2.1 WCE formalisation

Machine translation (MT) consists in finding the most probable target language sequence $\hat{e} = (e_1, e_2, ..., e_N)$ given a source language sentence $f = (f_1, f_2, ..., f_M)$.

We can represent Word-level Confidence Estimation (WCE) information as a sequence q (same length N of

*Toolkit name anonymized

[1]Since 2012 (http://www.statmt.org/wmt12/quality-estimation-task.html)

$\hat{e}$) where $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$[2].

Basically, the WCE component solves the equation:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p(q/f, e)\} \quad (1)$$

This is a sequence labelling task that can be solved with several machine learning techniques such as Conditional Random Fields (CRF) (Lafferty et al., 2001). However, to train sequence labelling models, we need a large amount of training data for which a triplet (f, e, q) is available.

2.2 Related work

According to Luong et al. (2015), features for WCE can be classified in two types regarding their origin: the "external features" and the "internal features". On one hand, internal features are extracted from the SMT system itself like alignment table, N-best list, word graph, *etc.* On the other hand, external features mainly come from linguistic knowledge sources like syntactic parser, WordNet or BabelNet API, *etc.* In our approach, we use both types of features. They are detailed in section 3 and 4.

The first work about confidence estimation (Ueffing et al., 2003; Blatz et al., 2004), focused at the word level, was inspired by work done in automatic speech recognition (Wessel et al., 2001). The combination of a large amount of features, through a Naive Bayes model and a Neural Network, showed that Word Posterior Probabilty (WPP) was the most relevant internal feature. Later on, Xiong et al. (2010) integrated POS tagging and other external features. In the same way, Felice and Specia (2012) proposed 70 linguistic features for quality estimation at sentence level. Some of theses features can be applied at word level. Their work also revealed the need of efficient machine learning algorithms to integrate multiple features and achieve better performance.

Recent workshops proposed a shared evaluation task of WCE systems, in which several attempts of participants to mix internal and external features were successul. The estimation of the confidence score uses mainly classifiers like Conditionnal Random Fields (Han et al., 2013) (Luong et al., 2014), Support Vector Machines (David et al., 2012) or Perceptron (Bicici, 2013).

Further, some investigations were conducted to determine which feature seems to be the most relevant. David et al. (2012) proposed to filter features using a forward-backward algorithm to discard linearly correlated features. Using Boosting as learning algorithm, Luong et al. (2015) were able to take advantage of the most significant features.

Our work, inspired by all those previous papers, proposes to mix internal and external features and uses CRF as decision algorithm to estimate a WCE score. The technical novelty is their integration in a single toolkit, with ready-made scripts, to quickly run experiments on different language pairs.

[2] q_i could be also more than 2 labels, or even scores but this paper only deals with error detection (binary set of labels)

3 Available Features

Our toolkit extracts several internal and external features to train a classifier, as indicated in Table 1. Some of them are already described in detail in our previous paper (*reference hidden for anonymity*). Consequently, the novel features, which we added into our current toolkit, are in **bold** format in Table 1. Also, the features written in *"italic"* are conventional features but extracted using new approach compared to our previous work so we give a more detailed description of them in subsection 4.2.3.

This list could be extended (by us or by other contributors) in the future, since the toolkit is made available to the research community. For instance, we plan to integrate the use of monolingual or bilingual word embeddings following the works of Mikolov et al. (2013).

It is important to note that our toolkit extracts the features regarding *tokens* in the machine translation hypothesis sentence. In other words, one feature is extracted for each token in the MT output. So, in the table, *left* and *right* refer to a feature extracted from a token positionned on the left or right side of considered word. Similarly, *source* refers to a feature extracted from the source word aligned to the considered word. More details on some of these features are given in the next section.

4 Toolkit

In this section, we detail our toolkit, which is a complete out-of-the-box WCE system. It is a customizable, flexible, and portable platform to build a CE system.

4.1 Pipeline Overview

Our toolkit is described in Figure 1. It contains three essential components: *preprocessing, extracting features* and *training / decoding*. It integrates several existing NLP tools and API. It is developed in *Python 3* in order to use efficiently existing libraries/toolkits as well as object-oriented design.

The source code and ready-made scripts to run experiments on a French–English WCE task (for which the data is also made available) are given on a *github* repository (*link anonymized*).

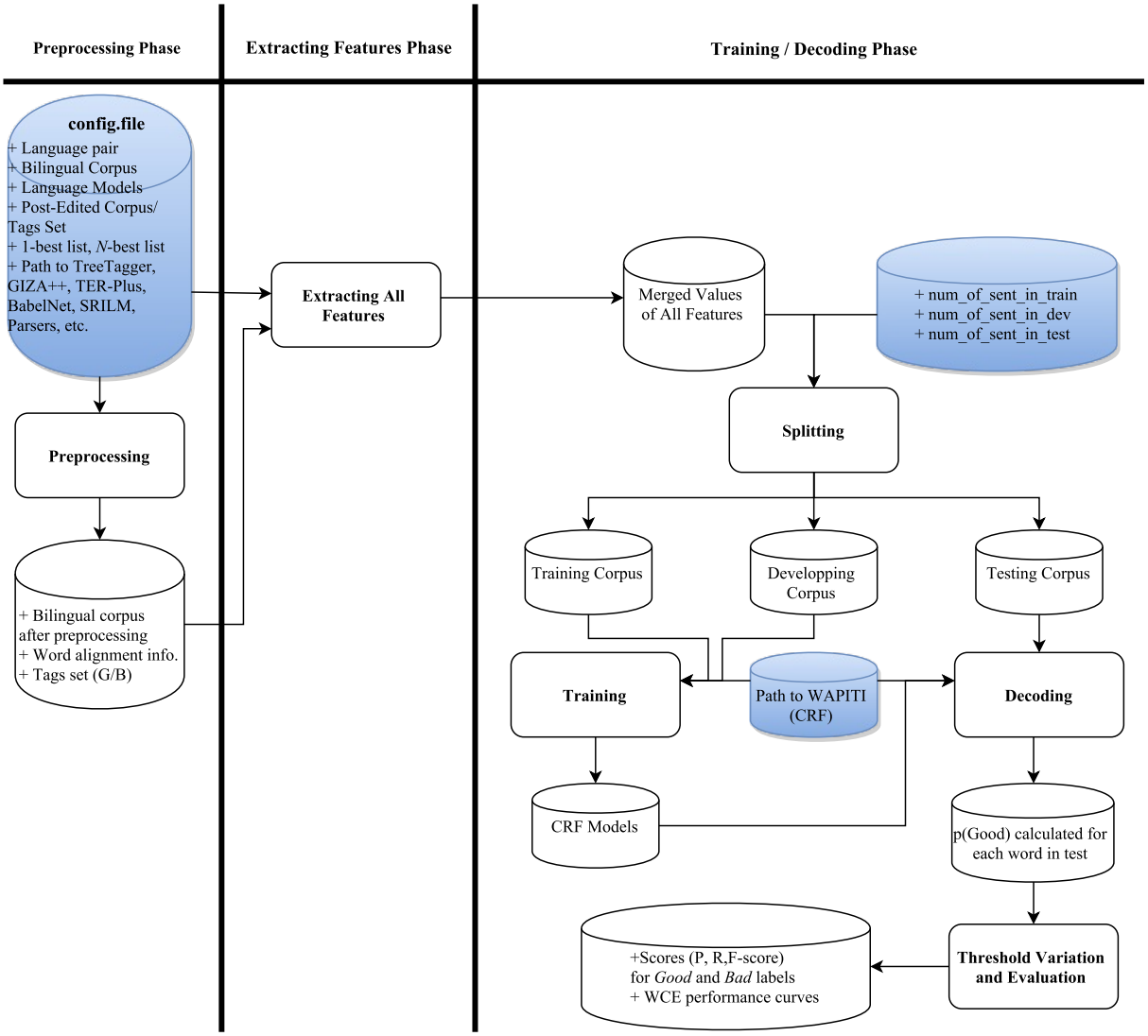
4.2 System Design

In this section, we shows different phases of our toolkit operations and usage of configuration parameters for French–English (fr–en) language-pair.

4.2.1 Configuration file

A configuration file gathers the main WCE parameters. It is stored in YAML[3] format. The configuration parameters are presented as following:

[3] `http://www.yaml.org/`



Note: above objects that have blue background are described in section 4.2.1

Figure 1: Pipeline of our word confidence estimation tool

1 *Proper Name*	9 *Left Source Word*	17 Left Target POS	25 WPP Exact	33 Punctuation
2 **Unknown Stemming**	10 **Left Source Stem**	18 Left Target Word	26 WPP Any	34 Stop Word
3 Number of Word Occurrences	11 Right Source POS	19 **Left Target Stem**	27 Max	35 Occur in Google Translate
4 **Number of Stemming Occurrences**	12 *Right Source Word*	20 Right Target POS	28 Min	36 **Occur in Bing Translator**
5 Source POS	13 **Right Source Stem**	21 Right Target Word	29 Nodes	37 *Polysemy Count – Target*
6 *Source Word*	14 Target POS	22 **Right Target Stem**	30 *Constituent Label*	38 Backoff Behaviour – Target
7 **Source Stem**	15 Target Word	23 Longest Target N-gram Length	31 *Distance To Root*	
8 Left Source POS	16 **Target Stem**	24 Longest Source N-gram Length	32 Numeric	

Table 1: Features extracted by the toolkit

source_language and **target_language**: specify the abbreviations of source and target language, respectively. Also, **language_pair** combines two above abbreviations. For instance, for language-pair fr–en, we have the following configuration parameters:

- source_language: fr
- target_language: en
- language_pair: fr_en

raw_corpus_src_lang, raw_corpus_tgt_lang : the file name of the given input corpus and output corpus after machine translation using, for instance, a phrase-based SMT system like *moses* (Koehn et al., 2007) ;

post_edition_file_path (name of the post-edited corpus if available - needed for WCE labels setting) or the **tags_file_path** (the WCE tags themselves if available *a priori*) - these labels are obviously needed for WCE training ;

lang_model_src_lang, lang_model_tgt_lang : the given language models (*arpa* format) for source and target languages, respectively ;

one_best_includ_alignment,
n_best_includ_alignment : the SMT output (in the format of *moses* (Koehn et al., 2007)) that consists in *1-* or N-best hypotheses of translations and including source-target word alignment information ;

num_of_sent_in_training,
num_of_sent_in_developping,
num_of_sent_in_testing : given number of sentences for training, development and testing corpus, respectively (parallel corpus will be segmented sequencially using these numbers) ;

google_translator_path, bing_translator_path : paths to output translations from Google Translator and Bing Translator of the same given source corpus *raw_corpus_src_lang* ;

Tools paths point out the paths to several toolkits such as:

- **tool_get_constituent_fr**: path of tool for obtaining French constituent tree (we currently use *BONSAI*[4]) ;
- **tool_berkeley_parser**: path of tool for obtaining constituent tree for several other languages (including English and Spanish) ;
- **tool_ngram**: path of language modelling toolkit (SRILM (Stolcke, 2002a) in this case) ;
- **tool_babel_net**: path to the script which calls BabelNet API[5] (for counting the number of senses of a given word);
- **tool_tree_tagger**: path to *TreeTagger* toolkit (Schmid, 1995) for POS extraction ;
- **tool_terpa**: path to *TER-Plus* toolkit (Snover et al., 2008) for WCE labels setting ;
- **tool_giza**: path to *GIZA++* toolkit (Och and Ney, 2003) for automatic word alignment ;
- **tool_wapiti**: path to *WAPITI* toolkit (Lavergne et al., 2010a) for CRF training and decoding ;

4.2.2 Preprocessing Phase

Pre-processing basically consists in obtaining POS tags, word alignments and all needed analyzes from the available parallel corpus (the target being a MT output). First, input data is lowercased and/or tokenized if necessary. Then, TreeTagger toolkit (Schmid, 1995) is applied to get the Part-Of-Speech (POS) tags and stem of each word in both source and target languages. Finally, word alignments are obtained using GIZA++ (Och and Ney, 2003). The different parsers are also applied on the data.

4.2.3 Features Extraction

Internal Features

- **Proper Name**: indicates if a word is a proper name or not ; named entity POS over the languages has to be normalized for this (for instance, the proper name POS for French is 'NAM' and for English they are 'NP, NPS').
- **Unknown Stemming**: the output of TreeTagger toolkit (Schmid, 1995) is processed to detect whether the stem of each word is known or not.
- **Number of Word Occurrences** and **Number of Stemming Occurrences**: count the occurrences of a word (or a stem) in the sentence.

[4] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html

[5] http://babelnet.org

- **Alignment context features**: these features (#5-22 in tab. 1) were initially proposed by (Nguyen et al., 2011) in regard with the intuition that collocation is a believable indicator for judging if a target word is generated by a particular source word. We also apply them in our experiments, containing:
  - *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.
  - *Target alignment context features*: the combinations of the source word and each word in the window ± 2 (one before, one after) of the target word.

  For instance, table 2 shows the hypothesis sentence *'the nature of the independence granted is also important .'* given its source sentence *'la nature de l' indépendance octroyée est aussi importante .'*. Therefore, in case of target word "of", the *source alignment context features* (window ± 1) are "of/nature", "of/de" and "of/l' "; while the *target alignment context features* (window ± 2) are "de/the", "de/nature", "de/of", "de/the" and "de/independence". Therefore, using bilingual word alignment information, our toolkit extracts 18 features (from 5 to 22 in table 1).

Target words	Source aligned words
the	la
nature	nature
of	de
the	l'
independence	indépendance
granted	octroyée
is	est
also	aussi
important	importante
.	.

Table 2: Example of parallel sentence to illustrate the extraction of alignment context features

- **Longest Target N-gram Length**: Applying SRILM toolkit (Stolcke, 2002b) for source and target language models permits to compute the length of the longest sequence created by the current word and its previous ones in the language model. For example, with the target word w_i: if the sequence $w_{i-2}w_{i-1}w_i$ appears in the target language model but the sequence $w_{i-3}w_{i-2}w_{i-1}w_i$ does not, the longest target n-gram value for w_i will be 3. The value set for each word hence ranges from 0 to 4 if we use 4-gram LMs.

- **Longest Source N-gram Length**: this feature is very similar to the previous one. Similarly, we compute the same value for the word aligned to w_i in the source language model using word alignment information.

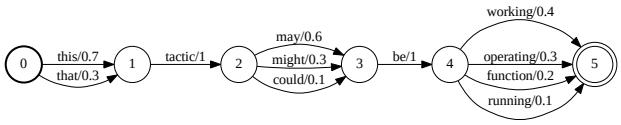


Figure 2: Example of Confusion Network

- **Word Posterior Probability** (WPP) and **Max**, **Min** and **Nodes** features: we build a confusion network from the SMT N-best-list using a dedicated tool. Then, the features are extracted: if we extract WPP of each word in the exact same position, we have feature **WPP Exact**. On the other hand, we have **WPP Any** when we calculate WPP of each word in any position. If we take the example of target word *"function"* that belongs to the confusion net example in Figure 2, its WPP is *0.2*; the number of alternative edges is *4*; the maximum and minimum values of the posterior probability distribution are *0.4* and *0.1*, respectively. It means that the graph topology features of word *"function"* are *Nodes = 4; Min = 0.1; Max = 0.4*.

External Features

- The word's constituent label (**Constituent Label**) and its depth in the constituent tree (**Distance to Root**) are also extracted (see example in Figure 3 which illustrates the distance between a word and its root in the tree). The constituent tree is obtained using Bonsai (for French) (Candito et al., 2010) or Berkeley parser (for English, Spanish) (Petrov and Klein, 2007).

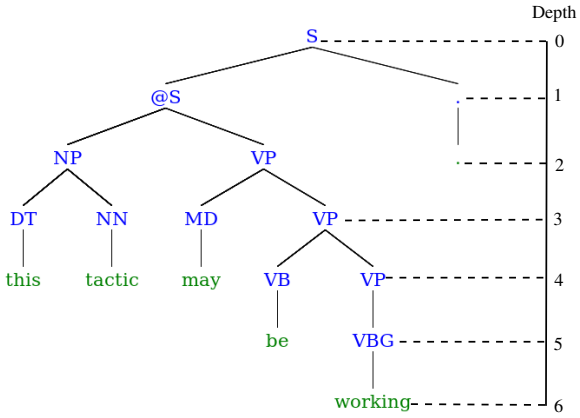


Figure 3: Example of Constituent Tree generated by Berkeley parser

For instance, after using toolkit Berkeley parser for English sentence *"this tactic may be working ."*, we have the following result: *((S (@S (NP (DT this) (NN tactic)) (VP (MD may) (VP (VB be) (VP (VBG working))))) (. .)))*. The figure 3 represents the constituent tree as well as the syntactic

Original Reference:	this	is	enough	to	shake	asset	prices	around	the	world	.	
Original Hypothesis:	what	is	enough	to	cower	prices	of	assets	around	the	world	.
Reference:	this	is	enough	to	*****	shake	asset	prices	around	the	world	.
Hypothesis:	what	is	enough	to	cower	prices	of	assets	around	the	world	.
Hypothesis After Shift:	what	is	enough	to	cower	of	assets	prices	around	the	world	.
Alignment:	S	E	E	E	I	S	T	E	E	E	E	E
HypErrs:	1.557	0.000	0.000	0.000	0.259	0.562	1.557	0.000	0.000	0.000	0.000	0.000

Table 3: Example of the Terplus toolkit's output processed

structure. Furthermore, based on the tree, we can obtain the constituent label and distance to root for each token word. In the case of *"working"*, these values are *VBG* and *6*, respectively.

In order to represent hierarchical structures and extract the two features, Natural Language ToolKit (NLTK) (Bird et al., 2009) is used.

- **Numeric**: we check whether this target word is numeric or not.

- **Punctuation, Stop Word**: we verify whether the target word appears in a given list of punctuations and in a given list of stop words, respectively.

- **Occur in Google Translate** and **Occur in Bing Translator**: in the translation hypothesis, we test the presence of the target word in on-line translations given respectively by *Google Translate* and *Bing Translator*.

- **Polysemy Count – Target**: to extract the polysemy count, we use the output of a BabelNet (Navigli and Ponzetto, 2012) API. BabelNet includes a network of semantic relations and allows us to extract the number of meanings of a word in a given language.

- **Backoff Behaviour – Target**: Our toolkit extracts how many times each word in the target sentence has to back off to assign a probalibity to its sentence. We use this information as a feature as done by Raybaud et al. (2011).

4.2.4 Estimating Phase

Once the final feature extraction stage has been completed, our scripts call Wapiti toolkit (Lavergne et al., 2010a) for training (or decoding) CRF models.

5 WCE Experiments

This section presents the experiments done for 2 different language pairs: French–English (*fr–en*) with the corpus provided by (Potet et al., 2012) and English–Spanish (*en–sp*) corresponding to the WMT shared task on word confidence estimation (2014[6] edition).

[6] http://www.statmt.org/wmt14/quality-estimation-task.html

5.1 The French–English post-edited corpus

The *fr–en* corpus contains 10881 translations. It was taken from several French–English news corpora from former WMT evaluation campaigns (from 2006 to 2010) (Potet et al., 2012).

To obtain the translations, Potet et al. (2012) used a French–English phrase-based translation system based on *moses* toolkit (Koehn et al., 2007). This medium-sized system was trained on Europarl and News parallel corpora for a former WMT evaluation shared-task (system more precisely described in (Potet et al., 2010) - 1.6M parallel sentences and 48M monolingual sentences in target language).

The hypotheses translated were post-edited according to the methodology described in (Potet et al., 2012). 10000 random sentences were extracted to create the training data and the remaining sentences were used for the evaluation corpus.

In order to evaluate our WCE system, we obtained a sequence q of quality labels (recall that $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$) using TER-Plus toolkit (Snover et al., 2008). Each word or phrase in the hypothesis e_{hyp} is aligned to a word or phrase in the reference (e_{ref}) with different types of edit: I (insertions), S (substitutions), T (stem matches), Y (synonym matches), P (phrasal substitutions) and "E" (exact match). Then, we re-categorize the obtained 6-label set into binary set: the E, T and Y belong to the $good$, whereas the S, P and I belong to the bad category.

An example of output of TER-Plus evaluation tool is shown in table 3.

5.2 Adaptation to a new language pair

To evaluate our toolkit on another language pair (en–es), we used the official data from WMT 2014 shared task on WCE.

One of the strength of our toolkit is the easiness to adapt it to another language pair within the (so-far) supported languages such as French, English, Spanish. Thus, a few configuration parameters showed in 4.2.1 were changed to move from the French–English to English–Spanish, like:

- source_language = en
- target_language = es
- language_pair = en_es
- raw_corpus_src_lang, raw_corpus_tgt_lang = given paths

		Our previous results (*ref. anonymized*)				Toolkit (default thr.)				Toolkit (oracle thr.)			
Task	**Label**	**P**	**R**	**F**	**M-F**	**P**	**R**	**F**	**M-F**	**P**	**R**	**F**	**M-F**
fr–en	*Correct*	85.99	88.18	87.07	62.41	84.19	**90.71**	**87.33**	**64.51**	85.60	85.65	85.62	**65.59**
	Incorrect	40.48	35.39	37.76		**50.31**	**35.59**	**41.69**		**45.61**	**45.50**	**45.56**	
en–es (WMT2014)	*Correct*	67.88	82.92	74.65	59.37	**71.24**	77.73	74.35	**60.76**	**71.42**	76.82	74.03	**60.87**
	Incorrect	54.22	37.18	44.10		51.82	**43.28**	**47.17**		51.49	**44.45**	**47.71**	

Table 4: The toolkit's WCE performances with fr–en and en–es language pairs. Note that the first set of results corresponds to the baseline work we did for WMT 2014. The second set of results is our toolkit performance with default decision threshold and the last set proposes some results with an oracle threshold (through mean F-measure of "correct" and "incorrect" labels).

- post_edition_file_path = *none* (for WMT 2014, the labels were directly given by the organizers of the shared task)
- tags_file_path = given path
- lang_model_src_lang, lang_model_tgt_lang = given paths
- one_best_list_included_alignment = n_best_list_included_alignment = *none* (N-best list was not provided for WMT shared task)

Consequently, our WCE toolkit executes en–es task in the same way as for fr–en task, but some features may not be extracted due to language-pair specificities: unavailable tools, no N-best, etc. For instance, for en–es, due to un-availability of N-best list, we could not extract the five following features: *WPP Exact, WPP Any, Nodes, Min, Max*.

5.3 Results

The WCE evaluation measures are the Precision (P), the Recall (R) and the F-score (F) of each label (as reminder, the decision label can be either "good" or "bad").

We use *wapiti* (Lavergne et al., 2010b) to train CRF model. The parameters used are the following: size of the interval for stopping criterion = 0.00005 ; maximum number of iterations done by the training algorithm = 200 and window size for the stopping criterion = 6 (options *-e 0.00005 -i 200 -w 6*).

Table 4 shows our classification performances for the French–English (*fr–en*) shared task and for the adaptation to a new shared task: English–Spanish (*en–es*). We present the Precision, Recall and F-Measure scores associated to both *good* (correct) and *bad* (incorrect) labels. These results are compared with State-of-the-Art results we previously obtained for WMT 2014 (*reference anonymized* - system ranked high in both 2013 and 2014 shared tasks)

The overall results (mean *F-mesure*) show that our toolkit, while integrating feature extraction in a single script, obtained similar performances (if not better) compared to our previous system. Using the default threshold of our classifier, the *fr–en* results are improved, specially the scores associated to the *incorrect* label. The *fr–en* F-measure associated to the *incorrect* label increased (+3.93 points) even if the precision is not improved. Due to the decrease of the recall, the *correct* label F-measure is slightly increased (+0.26 points).

Optimizing the threshold (oracle experiments since we do not have a development corpus), we strongly improve the scores for the incorrect label (+7.80 points) regarding the baseline in the *fr–en* task.

Also, the adaptation to the task *en–es* also obtained better results according to the mean F-Measure. The score of our toolkit with the default threshold improved the baseline of 1.39 points. If the oracle threshold is used, the improvement is 1.5 points, compared to the baseline.

The main improvements are obtained with the Precision of the *correct* label (respectively 3.36 and 3.54 points), while the Recall goes from 82.92 to 77.73 and 76.82, respectively. The resulting F-Measure of the *correct* label slightly decreases. Concerning the *incorrect* label, the Recall score is improved from 37.18 to 43.28 and 47.71. While the Precision decreases of 2.4 and 2.75 points. Overall, the F-Measure is improved for the default and oracle threshold.

6 Conclusions and Perspectives

This paper presented our Word Confidence Estimation (WCE) approach made available through a free toolkit. It combines some classical features and some new linguistic features. We use Conditional Random Fields as classifier to estimate the correctness of a word.

The WCE experiments conducted achieve state-of-the-art performance measures on two different data sets corresponding to two language pairs (French–English and English–Spanish).

Our feature extractors have been packaged into a free open-source toolkit in order to reproduce our experiments and allow researchers to train state-of-the-art WCE systems rapidly. This package is made available on a *github* repository (anonymous link).

We also believe that all the features engineering it includes could be used for other sequence labelling NLP tasks.

Further work will focus on adding features (word embeddings for instance) and language pairs. We also plan to extend our toolkit to the design of WCE for speech translation tasks.

References

Ergun Bicici. 2013. Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria, August. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

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

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *Proceedings of COLING'2010*.

Langlois David, Raybaud Sylvain, and Smaïli Kamel. 2012. Loria system for the wmt12 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 114–119, Baltimore, Maryland USA, June.

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

Aaron Li-Feng Han, Yi Lu, Derek F Wong, Lidia S Chao, Liangye He, and Junwen Xing. 2013. 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.

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. 2007. 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.

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

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

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

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2014. Lig system for word level qe task at wmt14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 335–341, Baltimore, Maryland USA, June.

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2015. Towards accurate predictors of word quality for machine translation: Lessons learned on french - english and english - spanish systems. *Data and Knowledge Engineering*, page 11, April.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Bach Nguyen, Fei Huang, and Yaser Al-Onaizan. 2011. 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.

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

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.

M Potet, L Besacier, and H Blanchon. 2010. The lig machine translation system for wmt 2010. In ACL Workshop, editor, *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*, Uppsala, Sweden, 11-17 July.

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. 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.

Sylvain Raybaud, David Langlois, and Kamel Smaili. 2011. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):p. 1–34, August.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

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

Andreas Stolcke. 2002a. Srilm - an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904.

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

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

Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*.

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