Strategy-Based Technology for Estimating MT Quality

Liugang Shang
Knowledge Engineering and Human-Computer Interaction center, Shenyang Aerospace University, Shenyang, China
1437260083@qq.com

Dongfeng Cai
Knowledge Engineering and Human-Computer Interaction center , Shenyang Aerospace University, Shenyang, China
caidf@vip.163.com

Duo Ji
Knowledge Engineering and Human-Computer Interaction center, Shenyang Aerospace University, Shenyang, China
jiduo_1@163.com

Abstract

This paper introduces our SAU-KERC system that achieved F1 score of 0.39 in the world-level quality estimation task in WMT2015. The goal is to assign each translated word a “OK” or “BAD” label indicating translation quality. We adopt the sequence labeling model, conditional random fields (CRF), to predict the labels. Since “BAD” labels are rare in the training and development sets, recognition rate of "BAD" is low. To solve this problem, we propose two strategies. One is to replace “OK” label with sub-labels to balance label distribution. The other is to reconstruct the training set to include more "BAD" words.

1 Introduction

QE task is proposed to estimate the quality of machine translation without relying on reference translations. It contains three levels -- word, sentence, and document and our work focuses on the word-level task. The word-level task was proposed in 2013 and was divided into binary classification and multi-class classification. This year only binary classification was considered in WMT2015.

OK/BAD: If a word need editing, then it is BAD. It is OK, otherwise.

As a confidence estimation problem, methods aim to confidence estimation before 2013. A lot of researchers started to investigate confidence measures for machine translation for nearly a decade (Gandrabur and Foster, 2003; Quirk, 2004; Ueffing et al., 2003). Many different confidence measures are investigated in(Blatz et al 2003). They are based on source and target language models features, n-best list, word-lattices, translation tables, and so on. The authors also present efficient ways of classifying words as ”correct” or “incorrect” by using native Bayes, single- or multi-layer perceptron. (Blatz et al 2003) combines several features and use neural network and naïve Bayes learning algorithms to predict whether a word is ok or bad. (Xiong et al., 2010) combines syntax feature, vocabulary feature and word posterior probability feature, which are extracted based on LG parsing, and use the binary classifier based on Maximum Entropy Model to predict the label of each word in machine translation(ok or bad).

Some good ideas are proposed in word-level QE task of WMT. (Luong et al., 2013) use both internal and external features into a conditional random fields(CRF) model to predict the label for each word in the MT hypothesis. (Wisniewskiet al., 2014) rely on a random forest classifier and 16 features to predict the label of a word. (Souza et al., 2014) train two classifier models by using bidirectional long short-term memory recurrent neural networks and CRF to complete word level QE Task.

In WMT2015, the high ratio of OK labels in the training set and development set makes the task an unbalanced classification problem. Generally, it is hard to solve unbalanced classification problem effectively using common machine learning algorithms and features. To balance the label distribution, we propose two strategies: refining OK label(ROL) and changing training set structure(CTS). We augment the CRF model with these two strategies to improve the performance.

The rest of this paper is organized as follows. Section 2 gives the selected features. Section 3 introduces the learning algorithm and the strategies we used. Section 4 shows the structure of experimental data. Section 5 analyzes the exper-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 348–352,
Lisboa, Portugal, 17-18 September 2015.

iment results. The last part is our summary of this task.

2 Feature

The features used in this paper were from portion of features provided by organizer and portion of (Luong et al., 2014) features.

2.1 Organizer's Feature

Target word: the combinations of target words in the window ±2(two before, two after of current word).
First aligned word: source word with maximum alignment probability with target word.
Is stop word: whether the target word is a stop word, punctuation symbol, proper name or number.
Back-off: 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).
Target/source pos: the target word pos and the source word pos; the bigram and trigram sequences.
Polysemy count: the number of senses of each word.

2.2 LIG System Feature

Target pos /target LM: the longest target word n-gram length and the longest target pos n-gram length.
Is in google: taking google translation as a pseudo-reference translation, we check whether a target word appear in the sentence generated by Google.

2.3 Other Feature

Target word frequency: the number of times the word appears in the machine translation result.
The distance between source and target word: the distance between positions of a target word and its aligned word in the sentence; if a target has not aligned word, then the distance is maximum.

2.4 Feature selection

In the CRF feature template, we chose 85 combinations of features in total. In fact, there are thousands of combinations of features which can be extended by the ten basic features, but too many features combined together do not contributed to the MT estimation system, instead this will cause a negative impact. Another problem is that if too much features are combined together, the current data set will have a good effect, but if the data set will appear for a bad effect, which is characterized by over-fitting. Thus feature selection is very critical for each system, and it directly affects the classifier accuracy and generalization ability.

At present, (Yu S H et al. 2007) feature selection can be divided into three strategies according to the formation of features subsets, namely global optimization, random search and heuristic search. Global optimization strategy commonly uses branch and bound algorithm, which search space is $O(2^n)$, random search strategy commonly use a genetic algorithm, which search space is smaller than $O(2^n)$. Heuristic search strategy commonly uses algorithms which have separate feature combination, the sequence former selection method (SFS), the sequence behind selection algorithms (SBS). Its search space is $O(N^2)$, although the heuristic search strategy has high efficiency, the result of heuristic search is not the global optimum(Yao Xu et al. 2012).

The selection method used in this paper is to add a feature to see if it has a contribution to the system. Eventually we keep 85 features, but it is not the optimal combination. We test data sets by using ten-fold cross-validation approach to prevent overfitting.

3 Labeling Method

Word level QE task of WMT2015 aims at marking each word in MT as OK or BAD. There must be some corresponding relationship among words in a MT output, so we also can regard word-level QE task as Sequence labeling task. We combine the ML method of CRF(using pocket CRF toolkit) with features describes in section 2 to train a sequence labeling model to predict word label.

The parameterization of CRF is shown as follows:

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{i,k} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l (y_i, x, i)\right)$$

t_k is defined as characteristic function at the edge, called transfer features which depend on the current position and the previous one; s_l is defined as characteristic function at the node, called state characteristics which depend on the current position. The conditional probability of each tag sequence equals to the sum of state probability and transfer probability of input sequence.

In QE task, the ratio between OK and BAD roughly equals to 4:1, which is very unbalanced. So it leads to two phenomena as fellows: 1. the probability labeling OK is much larger than the probability labeling BAD. 2. The probability that transfer to OK is much larger than the probability that transfer to BAD in train corpus; which will result in model bias. So the performance of the model trained just by using CRF and features of section 2 is not satisfactory.

In order to solve the unbalanced problem of word label, we propose two strategies: 1. Refine OK label(ROL); 2. Change train set structure(CTS).

3.1 Refine OK Label

We divide OK into OK_B, OK_I, OK_E and OK. OK_B is the start of OK continuous sequence; OK_I is the middle section of OK continuous sequence; OK_E is the end of OK continuous sequence; OK indicates the discontinuous label of OK as shown in figure 1. ROL can reduce the probability that a word is marked as OK to a certain extent. When we regard each label of words as a state, we can draw that ROL can reduce the probability of transfer to OK and enhance the probability of transfer to BAD tags in each output.

Target :	Es	totalmente	gratuito	y	esas	cosas	!
Label :	OK	OK	OK	OK	BAD	BAD	OK
Refine label :	OK_B	OK_I	OK_I	OK_E	BAD	BAD	OK

Figure 1: Refine OK Label

3.2 Change Train Set Structure

Our first strategy smooths the ratio between labels by refining OK label. However, even with refining, the proportion of BAD is still much smaller than other labels. So the second strategy we proposed will raise the proportion of bad by changing the structure of train set.
Implementation of this strategy:

a. Calculated the proportion of bad in each MT sentence in train set
b. Delete MT sentence that has no BAD label in train set.
c. MT sentence that BAD ratio is greater than threshold K be added repeatedly into train set.

This strategy will reduce the number of OK and increase the number of BAD, consequently reducing the ratio between OK and BAD.

4 Experiment

4.1 Data

There is just one translation corpus from English to Spanish in word-level QE task of WMT2015. The detail information of corpus shows in table 1:

	EN-ES		
	Train	Dev	Test
Sentence	11271	1000	1817
Word	257548	23207	40899
OK : BAD	4.22 : 1	4.21 : 1	4.30 : 1

Table 1: Corpus structural information

As shown in table 1, the proportion of OK and BAD unbalanced, which will lead to an offset model. It needs strategies in section 3 to balance the ratio between OK and BAD. The train set after processing show in table 2:

Train set	**Pre-process**	**Post-process**
sentence	11271	14559
word	257548	311998
OK/BAD	4.22 : 1	1:6.9
OK_B/BAD	///	1:3.7
OK_I/BAD	///	1.3:1
OK_E/BAD	///	1:3.7
OK_ALL/BAD	**4.2:1**	**1.9:1**

Table 2: Training data information after change

4.2 Threshold K Determination

There is a threshold K in the strategy of changing training set structure. The size of threshold has influence on MT estimation performance, so we conducted a series of tests to analysis the size of K. Meaningful range of the threshold value of K should ensure reducing the proportion of OK and BAD. From table 1, the ratio between OK and BAD is 4.22/1, so we set threshold in range of [0.2,0.95] in experiment, its step size is 0.05. Experiments were carried out when OK label is not refined on the development set. The testing result is shown in table 3:

K value	F_BAD	F_OK	F_all
0.20	0.348	0.871	0.771
0.25	0.349	0.870	0.770
0.30	**0.350**	0.872	0.772
0.35	0.348	0.874	0.773
0.40	0.344	0.873	0.772
0.45	0.342	0.875	0.773
0.50	0.333	0.877	0.773
0.55	0.330	0.879	0.774
0.60	0.327	0.879	0.774
0.65	0.329	0.881	0.775
0.70	0.325	0.881	0.774
0.75	0.320	0.881	0.774
0.80	0.317	0.882	0.773
0.85	0.319	0.882	0.774
0.90	0.318	0.882	0.774
0.95	0.318	0.882	0.774

Table 3: Threshold experiment

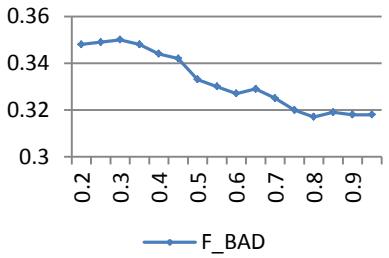

Figure 2: F score of BAD

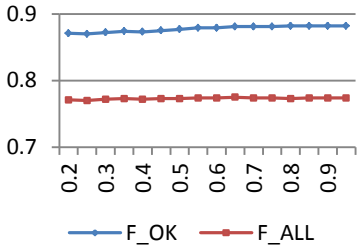

Figure 3: F score of OK

As shown in Figure2 and Figure3, changing in the threshold K have a certain effect on BAD label, but has little effect on the F1 score of OK and all labels. In Figure 1, the F1 score of BAD is highest when threshold K takes 0.3. However, we had set the value of K at 0.6 due to time reason during QE task. We believe that the score will be higher when K is equal to 0.3.

4.3 QE Experimental Analysis

There are four comparative experiments to prove the validity of the strategies proposed in this paper. Experiment names are as follows:

WY: do not change the structure of train set, not refine OK label.

WF: do not change the structure of train set, refine OK with OK_B, OK_I, OK_E, OK.

ZY: change the structure of train set, do not refine OK label.

ZF: change the structure of train set, refine OK label with OK_B, OK_I, OK_E, OK.

strategy	F_BAD	F_OK	F_AVG
WY	28.56	88.58	77.12
WF	34.53	87.63	77.44
ZY	32.71	88.16	77.52
ZF	38.34	86.84	77.53

Table 4: The results on development corpus

strategy	F_BAD	F_OK	F_AVG
WY	28.34	88.75	77.34
WF	34.28	87.97	77.83
ZY	32.69	88.3	77.80
ZF	39.11	86.36	77.44

Table 5: The results on test corpus

In QE task of WMT2015, Label distribution disequilibrium phenomenon can lead to Paranoid problem, which impacts the performance of QE system seriously. As shown in table 4 and table 5, the strategies that refine OK label and change structure of train set can solve label disequilibrium problem to a certain degree. The F_BAD is 34.28 when using the strategy of refining OK label alone, and the F_BAD is 32.69 when using the strategy of changing structure of training set. The strategy that refines OK label is more effective than the one that change the structure of the training set.

5 Conclusion

For the problem of Label distribution disequilibrium in word-level QE task of WMT2015, We proposed two strategies: one is refining OK label, the other one is changing structure of train set. Combined with the strategies, we use CRF and some grammar features to train a model which can enhance the correct number of BAD label, and the strategy of ROL is more effective. But, from Table 5, the F1 scores of the original method is that F_BAD is 28.34 and the F_OK is 88.75. When we add the two strategies, the F_BAD increases to 39.11 and the F_OK reduces to 86.36. In the future, we hope to overcome the shortcomings of the two strategies to improve both F1 scores of the two labels.

Reference

Blatz J, Fitzgerald E, Foster G, et al. Confidence estimation for machine translation[C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 315.

Quirk. 2004. Training a sentence-level machine translation confidence metric. In Proc. LREC, pages 825–828, Lisbon, Portugal, May.

Ueffing N, Macherey K, Ney H. Confidence measures for statistical machine translation[C]//In Proc. MT Summit IX. 2003.

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.

Xiong D, Zhang M, Li H. 2010. Error Detection for Statistical Machine Translation Using Linguistic Features[J]. Acl Proceedings of Annual Meeting of the Association for Computational Linguistics, 604-611.

Luong N Q, Lecouteux B, Besacier L, et al. LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. Proceedings of The eighth Workshop on Statistical Machine Translation, 2013:384--389.

Guillaume Wisniewski, Nicolas P écheux, et al. 2014. LIMSI Submission for WMT'14 QE Task. Proceedings of The ninth Workshop on Statistical Machine Translation, pages 348-354.

Jos é G. C. de Souza, Jes ús Gonz ález-Rubio, Christian Buck. 2014. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. Proceedings of The ninth Workshop on Statistical Machine Translation, pages 322-328

Ngoc-Quang Luong, Laurent Besacier, Benjamin Lecouteux. 2014. LIG System for Word Level QE task at WMT14. Proceedings of The ninth Workshop on Statistical Machine Translation,pages 335-341.

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

Yao Xu, Wang Xiaodan, Zhang Xi etc. Methods of feature selection [J]. Control and Decision, 2012,27(2);

Yu S H, Ma Z, Yang X H. 2007. Nonsmooth finite-time control ofuncertain second-order nonlinear systems[J]. J of Control Theory and Applications, 5(2): 171-176.