MT Tuning on RED: A Dependency-Based Evaluation Metric

Liangyou Li* **Hui Yu†** **Qun Liu*†**
* ADAPT Centre, School of Computing
Dublin City University, Ireland
† Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences, China
{liangyouli,qliu}@computing.dcu.ie
yuhui@ict.ac.cn

Abstract

In this paper, we describe our submission to WMT 2015 Tuning Task. We integrate a dependency-based MT evaluation metric, RED, to Moses and compare it with BLEU and METEOR in conjunction with two tuning methods: MERT and MIRA. Experiments are conducted using hierarchical phrase-based models on Czech–English and English–Czech tasks. Our results show that MIRA performs better than MERT in most cases. Using RED performs similarly to METEOR when tuning is performed using MIRA. We submit our system tuned by MIRA towards RED to WMT 2015. In human evaluations, we achieve the 1st rank in all 7 systems on the English–Czech task and 6/9 on the Czech–English task.

1 Introduction

Statistical Machine Translation (SMT) is modeled as a weighted combination of several features. Tuning in SMT refers to learning a set of optimized weights, which minimize a defined translation error on a tuning set. Typically, the error is measured by an automatic evaluation metric. Thanks to its simplicity and language independence, BLEU (Papineni et al., 2002) has served as the optimization objective since the 2000s. Although various lexical metrics, such as TER (Snover et al., 2006) and METEOR (Lavie and Denkowski, 2009) etc., have been proposed, none of them can truly replace BLEU in a phrase-based system (Cer et al., 2010).

However, BLEU has no proficiency to deal with synonyms, paraphrases, and syntactic equivalent etc. (Callison-Burch et al., 2006). In addition, as a lexical and n-gram-based metric, BLEU may be not suitable for optimization in a syntax-based model.

In this paper, we integrate a reference dependency-based MT evaluation metric, RED[1] (Yu et al., 2014), into the hierarchical phrase-based model (Chiang, 2005) in Moses (Koehn et al., 2007). In doing so, we explore whether a syntax-based translation system will perform better when it is optimized towards a syntax-based evaluation criteria. We compare RED with two other evaluation metrics, BLEU and METEOR (Section 2). Two tuning algorithms are used (Section 3). They are MERT (Och, 2003), MIRA (Cherry and Foster, 2012). Experiments are conducted on Czech–English and English–Czech translation (Section 4).

2 Evaluation Metrics

An evaluation metric, which has a higher correlation with human judgments, may be used to train a better system. In this paper, we compare three metrics: BLEU, METEOR, and RED.

2.1 BLEU

BLEU is the most widely used metric in SMT. It is lexical-based and language-independent. BLEU scores a hypothesis by combining n-gram precisions over reference translations with a length penalty.

A n-gram precision p_n is calculated separately for different n-gram lengths. BLEU combines these precisions using a geometric mean. The resulting score is subsequently scaled by a length penalty, which penalizes a hypothesis if it is shorter than references. Equation (1) shows a formula for calculating BLEU scores:

$$BLEU = BP \cdot \left(\prod_{n=1}^{N} p_n^{w_n} \right), \qquad (1)$$

where,

$$BP = \min\{1.0, \exp(1 - |r|/|h|)\},$$

[1] REference Dependency

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 428–433,
Lisboa, Portugal, 17-18 September 2015.

r and h are a reference and a hypothesis, respectively. In this paper, we use $N = 4$ and uniform weights $w_n = \frac{1}{N}$.

Even though widely used in SMT, BLEU has some pitfalls. Because of strictly relying on lexical sequences, BLEU cannot correctly score meaning equivalents, such as synonyms and paraphrases. It does not distinguish between content words and functional words as well. In addition, the penalty is not sufficient to be an equivalent replacement of n-gram recall.

2.2 METEOR

METEOR relies on unigrams but considers both precision and recall. It evaluates a hypothesis by aligning it to a reference. METEOR identifies all possible matches between a hypothesis-reference pair with the following matchers:

- **Exact**: match words that have the same word form.
- **Stem**: match words whose stems are identical.
- **Synonym**: match words when they are defined as synonyms in the WordNet database[2].
- **Paraphrase**: match a phrase pair when they are listed as paraphrases in a paraphrase table.

Typically, there is more than one possible alignment. In METEOR, a final alignment is obtained by beam search in the entire alignment space. Given the final alignment, METEOR calculates a unigram precision P and a unigram recall R by assigning different weights to function words and content words to distinguish them, as in Equation (2) and Equation (3).

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|} \quad (2)$$

$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|} \quad (3)$$

where m_i is the ith matcher, h_c and r_c are content words in a hypothesis and a reference, h_f and r_f are functions words in a hypothesis and a reference, respectively. Then the precision and recall are combined as in Equation (4).

$$Fmean = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \quad (4)$$

To consider differences in word order, a penalty is calculated on the basis of the total number (m) of matched words and the number (ch) of chunks. A chunk is defined as a sequence of matches, which are contiguous and have identical word order. The penalty is formulated as in Equation (5):

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}. \quad (5)$$

The final METEOR score is calculated as follows:

$$Score = (1 - Pen) \cdot Fmean. \quad (6)$$

α, β, γ, δ and w_i are constants, which can be optimized to maximize the correlation with human judgments.

By considering synonym, paraphrases, METEOR has shown to be highly correlated with human judgments. However, these resources are language-dependent. Besides, METEOR is unigram-based and thus has a lack of incorporating syntactic structures.

2.3 RED

Instead of collecting n-grams from word sequences as in BLEU, RED extracts n-grams according to a dependency structure of a reference, called *dep-ngrams*, which have two types: headword chain (Liu and Gildea, 2005) and fixed/floating structures (Shen et al., 2010). A headword chain is a sequence of words which corresponds to a path in a dependency tree, while a fixed/floating structure covers a sequence of contiguous words. Figure 1 shows an example of different types of dep-$n$grams.

A $Fmean$ score is separately calculated for each different dep-$n$gram lengths. Then, they are linearly combined as follows:

$$RED = \sum_{n=1}^{N} w_n \cdot Fmean_n \quad (7)$$

Inspired by other metrics, such as TERp (Snover et al., 2009) and METEOR, RED integrates some resources as follows:

- Stem and synonym: used to align words. This increases the possibility of matching a dep-$n$gram. Different matchers are assigned

[2] https://wordnet.princeton.edu/

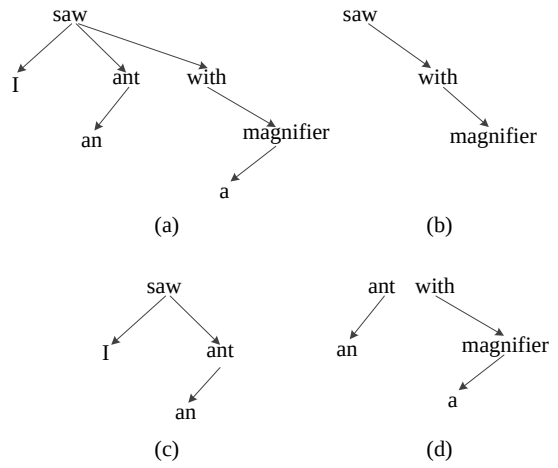

Figure 1: An illustration of dep-$n$grams. (a) is a dependency tree, (b) is a headword chain, (c) is a fixed structure and (d) is a floating structure.

different weights, this results in a scale factor for a dep-$n$gram as in Equation (8).

$$s_m = \frac{\sum_{i=1}^{n} w_{m_i}}{n} \tag{8}$$

- Paraphrase: used for extracting *paraphrase-ngrams*. In this case, RED ignores the dependency structure of a reference. A paraphrase-$n$gram has a weight w_{par}.

- Function Word: used to distinguish content words from function words. The function word score of a dep-$n$gram or a paraphrase-$n$gram can be calculated as follows:

$$s_f = \frac{cnt_f \cdot w_f + cnt_c \cdot (1 - w_f)}{cnt_f + cnt_c}, \tag{9}$$

where cnt_f and cnt_c are the number of function words and the number of content words.

Ideally, both a precision score P and a recall score R are based on the total number of dep-$n$grams in a hypothesis and a reference, respectively. However, in RED only dependency structures on the reference are available. Therefore, it uses the length of the hypothesis to approximate the number of the dep-$n$grams in the hypothesis to calculate P. Formulas for P and R are as follows:

$$P = \frac{score_{par} + score_{dep}}{|c|}, \tag{10}$$

$$R = \frac{score_{par} + score_{dep}}{Count_n(r) + Count_n(par)}, \tag{11}$$

where

$$score_{par} = \sum_{par \in P_n} w_{par} \cdot s_f, \tag{12}$$

$$score_{dep} = \sum_{d \in D_n} p(d, c) \cdot s_m \cdot s_f, \tag{13}$$

r and c are the reference and the hypothesis, P_n is the set of paraphrase-$n$grams, D_n is the set of dep-$n$grams. $p(d, c)$ is a match score which is 0 if no match is found; otherwise, it is a value between 0 and 1[3].

3 Tuning Algorithms

Tuning algorithms in SMT are designed to optimize decoding weights so that a defined translation error, typically measured by an automatic metric, is minimal on a development set. In this paper, we compare two algorithms: MERT and MIRA.

First, we introduce some notations. Let $\langle x, y \rangle \in D$ be a tuning set, where x and y are a source and a target, respectively. Let $\delta_y(d_x)$ be an error made by a derivation d on the source x given y as a reference. Let $\ell_m(D, \mathbf{w})$ be the total error measured by a metric m on the tuning set D with parameters $\mathbf{w}$.

3.1 MERT

MERT learns weights to rank candidate translations of each source sentence so that the final document-level score measured by a specific metric on the one-best translations is the highest. Formally, it tries to minimize the document-level error on the translations produced by the highest scoring translation derivation for each source sentence, as in Equation 14.

$$\ell_{MERT}(D, \mathbf{w}) = \oplus_{\langle x,y \rangle \in D} \delta_y(d_x^*), \tag{14}$$

where

$$d_x^* = \underset{d_x}{\text{argmax}} \ \mathbf{w} \cdot \Phi(d_x), \tag{15}$$

Φ are feature functions of the decoding model, $\mathbf{w} \cdot \Phi(d_x)$ is a score assigned to a deviation d_x

[3] If a headword chain $n$gram d in a reference r has a match in a hypothesis c, $p(d, c) = exp\{-\frac{\sum_{i=1}^{n-1} dist_{r_i} - dist_{c_i}}{n-1}\}$, where $dist_{r_i}$ and $dist_{c_i}$ are relative distances between ith word and $(i+1)$th word in the reference and hypothesis, respectively. If a fixed/floating structure is matched, $p(d, c) = 1$.

by the decoding model, $\oplus$ represents the accumulation of potentially non-decomposable sentential errors, which then produces a document-level evaluation score.

3.2 MIRA

MIRA is an online large margin learning algorithm (Crammer and Singer, 2003). Its application to MT decoding model tuning was firstly explored by Watanabe et al. (2007) and then refined by Chiang et al. (2008) and Cherry and Foster (2012). The MIRA we use tries to separate a "fear" derivation $d^-(x,y)$ from a "hope" one $d^+(x,y)$ by a margin propositional to their metric difference (Chiang et al., 2008). The two derivations are defined as follows:

$$d^+(x,y) = \operatorname*{argmax}_d \mathbf{w} \cdot \Phi(d) - \delta_y(d) \quad (16)$$

$$d^-(x,y) = \operatorname*{argmax}_d \mathbf{w} \cdot \Phi(d) + \delta_y(d) \quad (17)$$

Their model-score difference and metric-score difference are defined in Equation (18) and Equation (19), respectively.

$$\Delta s(x,y) = \delta_y(d^+(x,y)) - \delta_y(d^-(x,y)) \quad (18)$$

$$\Delta m(x,y) = \mathbf{w} \cdot (\Phi(d^+(x,y)) - \Phi(d^-(x,y))\} \quad (19)$$

Cherry and Foster (2012) adapt a batch strategy in MIRA. The error, that batch MIRA tries to minimize is defined as below:

$$\ell_{MIRA}(D,\mathbf{w}) = \frac{1}{2C}||\mathbf{w} - \mathbf{w}_0|| + \sum_{\langle x,y\rangle \in D} L(x,y) \quad (20)$$

where C is a constant and $L(x,y)$ is a loss over a source x and a reference y, which is defined in Equation (21).

$$L(x,y) = \max\{0, \Delta s(x,y) - \Delta m(x,y)\} \quad (21)$$

4 Experiments

We conduct experiments on Czech–English and English–Czech hierarchical phrase-based translation systems built using Moses with default configurations and default feature functions.

We use WMT newstest2014 as our development data, while our test data consists of the concatenation of newstest2012 and newstest2013, which includes 6,003 sentence pairs in total[4]. English sentences are parsed into dependency structures by Stanford parser (Marneffe et al., 2006). Czech sentences are parsed by a Perl implementation[5] of the MST parser (McDonald et al., 2005).

Train \ Eval.		BLEU	METEOR	RED
MERT	BLEU	18.90	28.38	19.91
	METEOR	18.68	28.64	20.02
	RED	18.07	28.17	19.97
MIRA	BLEU	19.12	28.54	20.02
	METEOR	19.10	28.56	20.05
	RED	17.74	28.82	20.02

Table 1: Czech–English evaluation performance. In each column, the intensity of shades indicates the rank of values.

4.1 Metrics Setting

As described in Section 2.1, we use the standard BLEU parameters[6]. We use METEOR 1.4[7] in our experiments with default optimized parameters. Specifically, for Czech to English translation, we adopt all four lexical matching strategies with parameter values: $\alpha = 0.85$, $\beta = 0.2$, $\gamma = 0.6$, $\delta = 0.75$ and $w_i = 1.0, 0.6, 0.8, 0.6$. For English to Czech translation, we use two lexical matching strategies, including *exact* and *paraphrase*, with parameter values: $\alpha = 0.95$, $\beta = 0.2$, $\gamma = 0.6$, $\delta = 0.8$ and $w_i = 1.0, 0.4$.

In RED, we use all four matchers in the Czech–English task while we do not use *stem* and *synonym* in the English–Czech task. The same parameter values are used in both tasks. We set $N = 3$, the corresponding $w_i = 0.6, 0.5, 0.1$. We set $w_{m_i} = 0.9, 0.6, 0.6$ for three matchers including exact, stem and synonym and $w_{par} = 0.6$ for the paraphrase matcher. We set $w_f = 0.2$ for function words and $\alpha = 0.9$ for combining P and R in $Fmean$.

4.2 Results

Table 1 and Table 2 show our experimental results on two tasks, respectively. We have several findings as below:

- In both tasks best scores are achieved when

[4] http://statmt.org/wmt14/translation-task.html

[5] http://search.cpan.org/~rur/Treex-Parser-MSTperl

[6] i.e., up to 4-gram matching with uniform weighting of n-gram precisions.

[7] http://www.cs.cmu.edu/~alavie/METEOR/

Train \ Eval.		BLEU	METEOR	RED
MERT	BLEU	11.25	17.36	14.95
	METEOR	10.44	17.00	14.86
	RED	9.51	16.81	14.58
MIRA	BLEU	11.52	17.54	15.14
	METEOR	11.43	17.56	15.26
	RED	11.29	17.67	15.25

Table 2: English–Czech evaluation performance. In each column, the intensity of shades indicates the rank of values.

MIRA is used rather than MERT. In most cases, MIRA is better than MERT.

- When RED is used in MERT, we obtain a worse performance than that of BLEU and METEOR in almost all cases, especially in the English–Czech task.
- When BLEU is used as the evaluation metric, the best score is obtained by using BLEU as the optimization objective in tuning as well. This follows the findings in Cer et al. (2010).
- The best METEOR score is achieved when RED is used to tune our system while the best RED score is obtained when METEOR is used to tune. Taking that the same resources are used in the two metrics into consideration, this may indicate that the two metrics are correlated.

5 Submission

We submit our system tuned by MIRA towards RED. In human evaluations, we get 6th out of 9 systems on the Czech–English task and the 1st rank in all 7 systems on the English–Czech task.

Such human judgments suggest that RED performs better on Czech than English. We guess this is because dependency n-grams have better capability of handling free word order in Czech sentences. This hypothesis can be an avenue for future work.

6 Conclusion

In this paper, we describe our submissions to WMT 2015 tuning task on Czech–English and English–Czech tasks. They are hierarchical phrase-based models both tuned by MIRA towards a dependency-based metric, RED. In human evaluations, our system gets the 1st rank in the English–Czech task.

Acknowledgements

This research has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

We thank Xiaofeng Wu for his discussion and anonymous reviewers for their insightful comments. In particular, we thank reviewer #2 for providing detailed suggestions.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The Best Lexical Metric for Phrase-based Statistical MT System Optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 555–563, Los Angeles, California.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Montreal, Canada.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online Large-margin Training of Syntactic and Structural Translation Features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, March.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi,

Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic.

Alon Lavie and Michael J. Denkowski. 2009. The Meteor Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, 23(2-3):105–115, September.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Language Resources and Evaluation*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 91–98, Ann Arbor, Michigan.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency Statistical Machine Translation. *Computational Linguistics*, 36(4):649–671, December.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER?: Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. 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 764–773, Prague, June.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. RED: A Reference Dependency Based MT Evaluation Metric. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2042–2051, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.