Blend: a Novel Combined MT Metric Based on Direct Assessment

— CASICT-DCU submission to WMT17 Metrics Task

Qingsong Ma[1] **Yvette Graham**[2] **Shugen Wang**[1] **Qun Liu**[2,1]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, University of Chinese Academy of Sciences
[2] ADAPT Centre, School of Computing, Dublin City University
maqingsong@ict.ac.cn, graham.yvette@gmail.com
wangshugen@ict.ac.cn, qun.liu@dcu.ie

Abstract

Existing metrics to evaluate the quality of Machine Translation hypotheses take different perspectives into account. DPMFcomb, a metric combining the merits of a range of metrics, achieved the best performance for evaluation of to-English language pairs in the previous two years of WMT Metrics Shared Tasks. This year, we submit a novel combined metric, Blend, to WMT17 Metrics task. Compared to DPMFcomb, Blend includes the following adaptations: i) We use DA human evaluation to guide the training process with a vast reduction in required training data, while still achieving improved performance when evaluated on WMT16 to-English language pairs; ii) We carry out experiments to explore the contribution of metrics incorporated in Blend, in order to find a trade-off between performance and efficiency.

1 Introduction

Automatic machine translation evaluation (AMTE) has received much attention in recent years, with the aim of providing quick and stable measurements of the performance of machine translation (MT) systems. Various metrics for AMTE have been proposed and most operate via computation of the similarity between the MT hypothesis and the reference translation. However, different metrics focus on different perspectives in terms of measuring similarity. For lexical based metrics, BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) count n-gram co-occurrence, Meteor (Denkowski and Lavie, 2014) and GTM (Melamed et al., 2003) catch different kinds of matches, ROUGE (Lin and Och, 2004) captures common subsequences, WER (Nießen et al., 2000), PER (Tillmann et al., 1997) and TER (Snover et al., 2009) compute the post-editing distance between the hypothesis and the reference translation. Syntactic based metrics mainly use shallow syntactic structures (Chan and Ng, 2008; Zhu et al., 2010), dependency tree structures or constituent tree structures (Owczarzak et al., 2007; Liu and Gildea, 2005). Semantic measures (Lo et al., 2012) and discourse similarity based metrics (Guzmán et al., 2014) have also been proposed.

Different metrics evaluate similarity between hypotheses and reference translations from various perspectives, each of which has pros and cons. One straightforward and effective method to take advantage of the merits of existing metrics is to combine quality scores assigned by these metrics, like DPMFcomb (Yu et al., 2015a).

In WMT15 and WMT16 Metrics tasks, DPMFcomb was the best metric on average for to-English language pairs (Stanojević et al., 2015; Bojar et al., 2016). DPMFcomb incorporates lexical, syntactic and semantic based metrics, using ranking SVM[1] to train parameters of each metric score and achieves a high correlation with human evaluation. Human evaluations in terms of relative ranking (RR) accumulated in WMT Metrics tasks are adopted to generate training data and to guide the training process. Human relative ranking is carried out by ranking the quality of 5 MT hypotheses of the same source segment from 1 to 5 via comparison with the reference translation.

[1]http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 598–603
Copenhagen, Denmark, September 711, 2017.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	en-ru
WMT15	500	500	500	–	500	–	500
WMT16	560	560	560	560	560	560	560

Table 1: The number of sampled DA data for each language pair in WMT15 and WMT16.

Therefore, human RR only provides relative differences in quality of a given 5 hypotheses rather than the overall absolute quality of hypotheses. Besides, the low inter-annotator agreement level in RR (Callison-Burch et al., 2007) has been a long-lasting issue in MT human evaluation. The ability and the reliability of RR raise our concern whether the capability of the model trained with RR as the golden standard may be limited.

Fortunately, a new emerged evaluation approach, direct assessment (DA) (Graham et al., 2013), has been proven more reliable for evaluation of metrics and was recently adopted as the official human evaluation in WMT17. DA produces absolute quality scores of hypotheses, by measuring to what extend the hypothesis adequately expresses the meaning of the reference translation, through a 1-100 continuous rating scale that facilitates reliable quality control of crowd-sourcing. Large numbers of repeat human assessments per translation are standardized and then combined into a mean score as the final quality score of the MT hypothesis.

The recent development in human evaluation of MT motivates us to propose a new combined metric, named as Blend [2], by adopting DA, as opposed to RR, to guide the training process indicating that a more reliable gold standard can lead to more reliable results even with less training data. Furthermore, we explore the contribution of metrics incorporated in Blend, aiming at finding a trade-off between performance and efficiency of Blend.

What follows is a brief review of DPMFcomb, before a description of Blend formulation is provided in Section 2, followed by experiments and results in Section 3, before the conclusions in section 4.

2 Metrics

2.1 Review of DPMFcomb

DPMFcomb utilizes human relative ranking data to train a combined metric that produces quality scores for MT hypotheses. In the training process, metrics are incorporated as features in the form of metric scores attributed to the same hypotheses, with relative ranks as the gold standard to guide SVM-rank to learn parameters for features. When testing, the predicted ranking scores produced by DPMFcomb reflect the quality of hypotheses. DPMFcomb allows the combination of the advantages of a set of arbitrary metrics resulting in a metric with a high correlation with human assessment. DPMFcomb includes default metrics provided by Asiya MT evaluation toolkit (Giménez and Màrquez, 2010), as well as three other metrics, namely ENTF (Yu et al., 2015c), REDp (Yu et al., 2014) and DPMF (Yu et al., 2015b). Over the past two years of WMT metrics tasks, DPMFcomb has achieved the best performance for evaluation of MT of to-English language pairs.

2.2 Blend: A Novel Combined Metric based on DA

Although RR reflects the quality of hypotheses to some extent, it has two obvious defects. Firstly, RR provides relative ranks of the given competing MT hypotheses, which only reflects relative differences in quality rather than the absolute quality of hypotheses. On the other hand, RR suffers from low inter-annotator agreement levels. As a result, the capability of the model trained with RR as the golden standard could be limited. However, DA with carefully design of criteria (Graham et al., 2013) produces highly reliable overall quality scores for each hypothesis (Graham et al., 2015). In addition, since DA has replaced RR as the official human evaluation in the news domain in WMT17, more DA data would become available in the coming years. These motivate our new combined metric, specially designed based on DA, rather than RR, named as Blend, which means it is a metric that can blend advantages of arbitrary metrics in a combined metric that has a high correlation with human assessment.

Our metric follows the basic formulation of DPMFcomb. However, since DA is an absolute quality judgment, which is different from RR, the

[2]Blend is available: https://github.com/qingsongma/blend

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg
Blend.all	**.991**	**.954**	.969	.879	**.942**	.972	**.951**
MPEDA	.988	.923	**.971**	**.905**	.923	**.975**	.948
BEER	.985	.871	.964	.828	.894	**.975**	.920

Table 2: System-level Pearson correlation of metric scores and DA human scores with 10K hybrid systems for to-English language pairs on WMT16, where "avg" denotes the average Pearson correlation of all language pairs.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg
Blend.all	.710	**.615**	**.602**	.636	**.622**	.658	**.641**
DPMFcomb	**.713**	.598	.584	.627	.615	**.663**	.633
METRICS-F	.696	.601	.557	**.662**	.618	.649	.631

Table 3: Segment-level Pearson correlation of metric scores and DA human scores for to-English language pairs on WMT16, where "avg" denotes the average Pearson correlation of all language pairs.

training data and the method of Blend are different from that of DPMFcomb. We employ SVM regression from libsvm (Chang and Lin, 2011) [3] for training, with training data consisting of features in terms of incorporated metric scores for hypotheses and the gold standard in terms of DA human scores.

3 Experiments

We carry out experiments to compare the performance of DPMFcomb and Blend. We also explore the contribution of incorporated metrics in Blend to find a trade-off between performance and efficiency.

3.1 Setups

Our experiments are tested on WMT16 to-English and English-Russian (en-ru) language pairs. We use DA data sampled from WMT15 and WMT16 (Table 1) for Blend. Since there is only a limited amount of DA data available at present, we employ all other to-English DA data as training data (4800 sentences) when testing on each to-English language pair (560 sentences) in WMT16. For en-ru, we use en-ru DA data in WMT15 (500 sentences) to train and test on en-ru DA data in WMT16 (560 sentences).

Features in both the training data and the test data are scaled to be in [-1,1]. We use epsilon-SVR with RBF kernel, and the epsilon is set to 0.1.

[3] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

3.2 Blend vs DPMFcomb

In WMT16, DPMFcomb incorporates 57 metrics and was trained with SVM-rank on 445K training segments extracted from WMT12-WMT14 to-English language pairs according to human judgments in terms of RR. For comparison, Blend incorporates the same 57 metrics but is trained with SVM regression on only 4,800 training data extracted from sampled DA data in WMT15-WMT16 for each to-English language pair. We name it Blend.all.

We present the system and segment-level Pearson correlation results in Table 2 and Table 3, respectively. Table 2 shows Blend.all has higher average system-level Pearson correlation (.951) with DA human scores compared to the two high performing metrics MPEDA (.948) and BEER (.920) on WMT16 for to-English language pairs.

Table 3 shows segment-level Pearson correlations of Blend.all and two other high-performing metrics DPMFcomb and EMTRICS-F on WMT16 for to-English language pairs. From Table 3 we can see Blend.all achieves the best performance in 3 out of 6 to-English languages pairs and state-of-the-art performance on average. It is worth noting that even though the training data of Blend.all is far less than that of DPMFcomb, Blend.all has higher average Pearson correlation (.641), trained on DA scores, than that of DPMFcomb (.633), trained on RR scores.

In all, the above results show Blend trained with DA data outperforms DPMFcomb trained with RR data on WMT16 for to-English language pairs.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg
Blend.all	**.710**	**.615**	**.602**	**.636**	**.622**	.658	**.641**
Blend.lex	.704	.589	.583	.625	.620	**.674**	.632
Blend.syn	.656	.528	.494	.560	.533	.610	.564
Blend.sem	.610	.533	.492	.507	.501	.554	.533

Table 4: Segment-level Pearson correlation of Blend incorporating different level of linguistic metrics for to-English language pairs on WMT16, where "avg" denotes the average Pearson correlation of all language pairs.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg
Blend.lex	.704	.589	.583	.625	.620	.674	.632
Blend.lex+CharacTer	.707	.596	.575	.628	.620	**.680**	.634
Blend.lex+BEER	**.709**	.589	.580	.627	.622	.673	.634
Blend.lex+DPMF	.706	.592	**.590**	.632	.626	.670	.636
Blend.lex+ENTF	.703	.595	.588	.629	.629	.676	.637
Blend.lex+4	**.709**	**.601**	.584	**.636**	**.633**	.675	**.640**

Table 5: Segment-level Pearson correlation of Blend.lex incorporating 4 other metrics for to-English language pairs on WMT16, where "avg" denotes the average Pearson correlation of all language pairs.

3.3 Trade-off between Performance and Efficiency

It is convenient for Blend to combine arbitrary metrics in order to achieve a high correlation with human assessment. However, it would be useful to know if any metric does not contribute greatly to Blend in terms of performance, while at the same time leads to low efficiency. To explore this, we separate out the default metrics for to-English language pairs provided by Asiya toolkit into three categories, namely, lexical, syntactic, and semantic based metrics. Blend.lex is the variant that incorporates only default lexical based metrics in Asiya toolkit, while Blend.syn, and Blend.sem. incorporate only syntactic and semantic metrics, respectively. Blend.lex includes 25 metrics, but with only 9 kinds of metrics, since some of them are simply different variants of the same metric. Blend.syn includes 17 metrics and Blend.sem 13 metrics but in reality each only corresponds to 3 distinct metrics, similar to Blend.lex.

The experimental results on WMT16 are shown in Table 4. It is not all that surprising that Blend.all incorporated with all default Asiya metrics achieves the best performance in 5 out of 6 language pairs and on average. However, it may be worth noting that the average Pearson correlation of Blend.lex is only 0.009 less than that of Blend.all, while the performance of Blend.syn and Blend.sem are quite far worse than that of Blend.all, and even that of Blend.lex. Since syntactic and semantic based metrics are usually complex, and the performance of Blend.lex is comparable with that of Blend.all, Blend can operate effectively with only incorporating the default lexical based metrics from Asiya toolkit.

We further add 4 other metrics to Blend.lex., CharacTer(Wang et al., 2016), a novel character-based metric; BEER(Stanojević and Sima'an, 2015), a metric combining different kinds of features; DPMF and ENTF, which proved to be effective. All of these 4 metrics are convenient to use. Table 5 shows *Blend.lex+4* (.640) achieves better performance than that of Blend.lex (.632), and is very close to that of Blend.all (.641) as shown in Table 3.

Hence, we submit *Blend.lex+4* to WMT17 Metrics task for to-English language pairs, since it provides a good trade-off between performance and efficiency for Blend.

3.4 Experiments on from-English language pairs

Blend can be effective to evaluate the quality of from-English MT hypotheses if incorporated metrics support from-English language pairs. We carry out experiments on WMT16 for en-ru language pair as shown in Table 6.[4] Blend.default

[4]For from-English language pairs, there is only en-ru DA data available at present.

	en-ru
Blend.default	.613
Blend.default+2	**.675**
BEER	.666

Table 6: Segment-level Pearson correlation for en-ru in WMT16.

is trained on only 500 sentences and incorporates default lexical based metrics from Asiya toolkit for en-ru, including 20 metrics, but with 9 kinds of metrics only. Compared with Blend.default, Blend.default+2 incorporates two more metrics, CharacTer and BEER, but achieves great improvement with segment-level Pearson correlation from .613 to .675. The incorporated metric BEER is the best performing metric (.666) on WMT16 for en-ru, which is trained with large amounts of data. Beer contributes to Blend apparently, meanwhile Blend can further improve the performance of BEER, indicating the effectiveness of the combined metric Blend. We submit Blend.default+2 to WMT17 Metrics task for en-ru.

4 Conclusions

The performance of DPMFcomb proves the effectiveness of the idea of combining metrics. However, DPMFcomb cannot extend itself to the new development of human evaluation. Therefore, we propose a novel metric Blend to employ DA data. Blend is also a combined metric that can take good advantage of the merits of existing metrics, and performs better than DPMFcomb, even with far less training data. Blend is easy to be trained and flexible to be applied to any language pairs. In this paper we present experiments on WMT16 Metrics task, which shows Blend achieves state-of-the-art performance on average for to-English language pairs and for en-ru. Furthermore, we carry out experiments with different settings and find a good trade-off for Blend in terms of performance and efficiency.

Acknowledgments

This research is supported by Chinas NSFC grant 61379086 and the European Union Horizon 2020 Programme (H2020) under grant agreement no. 645452 (QT21). 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.

References

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 199–231. http://www.aclweb.org/anthology/W16-2302.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 136–158.

Yee Seng Chan and Hwee Tou Ng. 2008. Maxsim: A maximum similarity metric for machine translation evaluation. In *ACL*. pages 55–62.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*. Citeseer.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., pages 138–145.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An open toolkit for automatic machine translation (meta-) evaluation. *Prague Bull. Math. Linguistics* 94:77–86.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *HLT-NAACL*. pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*. pages 33–41.

Francisco Guzmán, Shafiq R Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *ACL (1)*. pages 687–698.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 605.

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.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 243–252.

I Dan Melamed, Ryan Green, and Joseph P Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, pages 61–63.

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Evaluating machine translation with lfg dependencies. *Machine Translation* 21(2):95–119.

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*. Association for Computational Linguistics, pages 311–318.

Matthew Snover, Nitin Madnani, Bonnie 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*. Association for Computational Linguistics, Athens, Greece, pages 259–268.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 256–273. http://aclweb.org/anthology/W15-3031.

Miloš Stanojević and Khalil Sima'an. 2015. Beer 1.1: Illc uva submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 396–401. http://aclweb.org/anthology/W15-3050.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Eurospeech*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *ACL 2016 First Conference on Machine Translation, Berlin, Germany*.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015a. Casict-dcu participation in wmt2015 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 417–421.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015b. An automatic machine translation evaluation metric based on dependency parsing model. *arXiv preprint arXiv:1508.01996* .

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015c. Improve the evaluation of translation fluency by using entropy of matched subsegments. *arXiv preprint arXiv:1508.02225* .

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. Red: A reference dependency based mt evaluation metric. In *COLING*. volume 14, pages 2042–2051.

Junguo Zhu, Muyun Yang, Bo Wang, Sheng Li, and Tiejun Zhao. 2010. All in strings: a powerful string-based automatic mt evaluation metric with multiple granularities. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pages 1533–1540.