MEANT 2.0: Accurate semantic MT evaluation for any output language

Chi-kiu Lo
NRC-CNRC
Multilingual Text Processing
National Research Council Canada
1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
chikiu.lo@nrc-cnrc.gc.ca

Abstract

We describe a new version of MEANT, which participated in the metrics task of the Second Conference on Machine Translation (WMT 2017). MEANT 2.0 uses idf-weighted distributional ngram accuracy to determine the phrasal similarity of semantic role fillers and yields better correlations with human judgments of translation quality than earlier versions. The improved phrasal similarity enables a subversion of MEANT to accurately evaluate translation adequacy for any output language, even languages without an automatic semantic parser. Our results show that MEANT, which is a non-ensemble and untrained metric, consistently performs as well as the top participants in previous years - including ensemble and trained ones - across different output languages. We also present the timing statistics for MEANT for better estimation of the evaluation cost. MEANT 2.0 is open source and publicly available.[1]

1 Introduction

We introduce a new version of MEANT, which participated in evaluating MT systems for all language pairs in the metrics task of the Second Conference on Machine Translation (WMT 2017). MEANT 2.0 is a non-ensemble and untrained metric that only requires a monolingual corpus in the output language to build the word embeddings and an automatic shallow semantic parser to obtain the predicate-argument structure to evaluate MT systems for a language pair. We have also build a degraded subversion, MEANT 2.0 - nosrl, to evaluate MT systems for any output language by removing the dependency on semantic parsers for semantic role labeling (SRL) the reference and the machine translations. The correlation of MEANT with human judgments has been improved by using both inverse document frequency (idf) and distributional ngram accuracy within the phrasal similarity calculation: the former to weight the importance of each word for better adequacy, the latter to to account for word reordering for greater fluency. Our results show that MEANT consistently performs as well as the top participants in previous years across different output languages, including ensemble and trained participants. We also present the timing statistics that show the relatively low cost of running MEANT. This highly portable and open source semantic MT evaluation metric is a more accurate alternative to BLEU in evaluating translation quality for low-resource languages.

[1] http://chikiu-jackie-lo.org/home/index.php/meant

2 The family of MEANT

MEANT and its variants (Lo et al., 2015, 2014; Lo and Wu, 2011a) evaluate translation adequacy by measuring the similarity of the semantic frames and their role fillers between the human reference and machine translations. Figure 1 illustrates the concept of MEANT - the semantic roles and their fillers of the reference translation match more with those of the MT2 than with those of the MT1, therefore MT2 is a more adequate translation than MT1.

MEANT consistently outperforms the commonly used automatic MT evaluation metrics, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), CDER (Leusch et al., 2006) and WER in correlation with human adequacy judgment. It is relatively easy to port to other languages. In the full version of MEANT, it required only a monolingual corpus (for eval-

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 589–597
Copenhagen, Denmark, September 711, 2017. ©2017 Association for Computational Linguistics

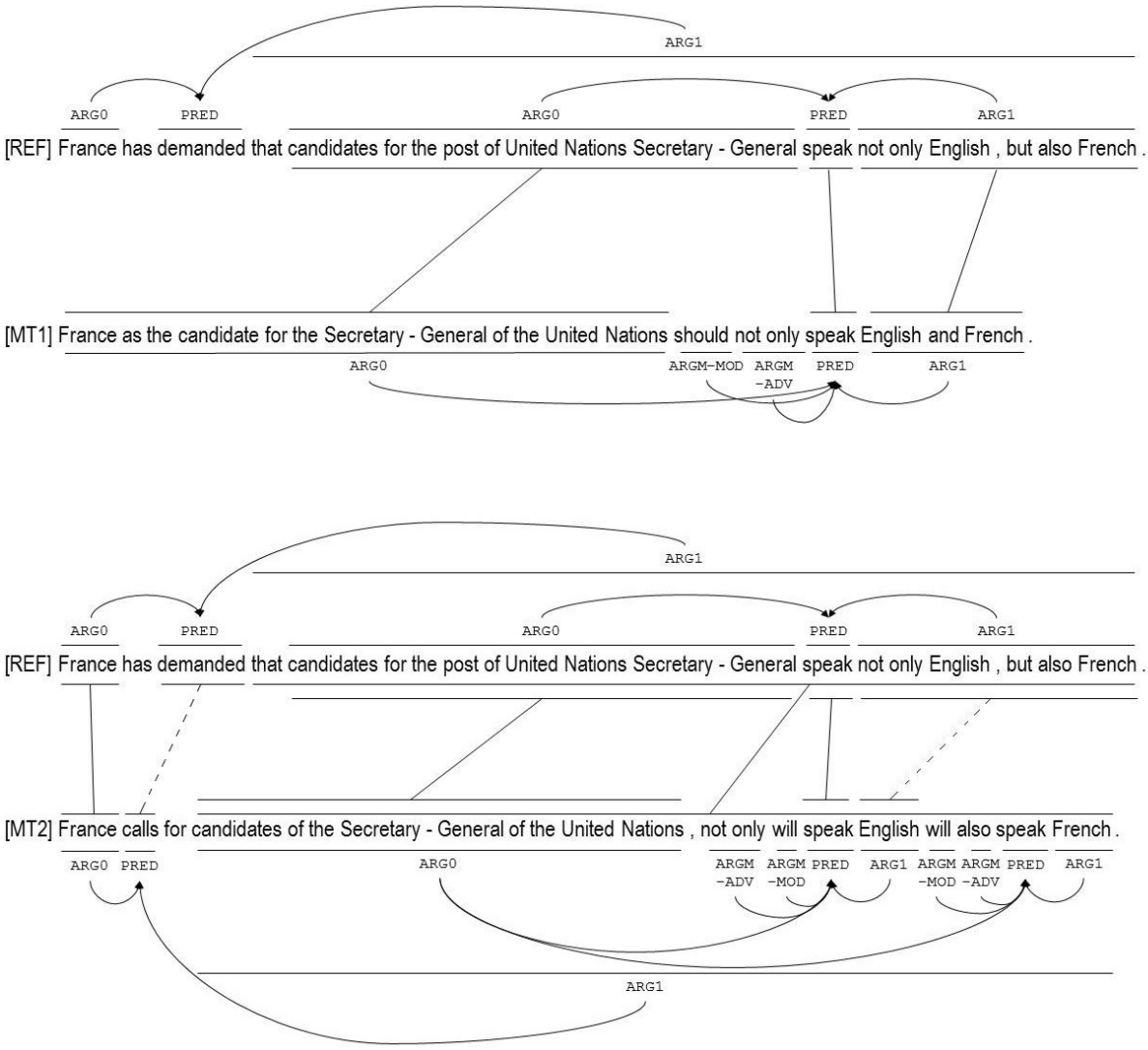

Figure 1: Example that illustrates the concept of MEANT. The solid role alignments mean the translation is mostly correct while the dotted role alignments mean the translation is partly correct. The semantic roles and fillers of the reference match more with those of MT2 than those of MT1, therefore MT2 is a more adequate translation than MT1.

uating lexical semantic similarity) and an automatic semantic parser (for evaluating frame semantic similarity) of the output language. In section 3, we describe a new subversion of MEANT that can be computed even when a semantic parser for the output language is unavailable.

MEANT is the weighted f-scores over corresponding semantic frames and role fillers in the reference and the machine translations. MEANT is generally computed as follows:

1. Apply a shallow semantic parser to both the reference and machine translations.

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the reference and machine translations according to the lexical similarities of the predicates.

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and MT output according to the lexical similarity of role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the following definitions:

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in MT} \\
q^1_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of REF}}{\text{total \#tokens in REF}} \\
w^0_{\text{nf}} &\equiv \frac{\text{\#tokens that are not fillers of any role inMT}}{\text{total \#tokens in MT}} \\
w^1_{\text{nf}} &\equiv \frac{\text{\#tokens that are not fillers of any role in REF}}{\text{total \#tokens in REF}} \\
w_{\{\text{pred}|j\}} &\equiv \text{weight of similarity of predicates or ARG } j \\
\mathbf{e}_{\text{sent}} &\equiv \text{the whole sentence string of MT} \\
\mathbf{f}_{\text{sent}} &\equiv \text{the whole sentence string of REF} \\
\mathbf{e}_{i,\{\text{pred}|j\}} &\equiv \text{role fillers of pred or ARG } j \text{ of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,\{\text{pred}|j\}} &\equiv \text{role fillers of pred or ARG } j \text{ of the aligned frame } i \text{ of REF} \\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f
\end{aligned}
$$

$$
\text{prec}_{\mathbf{e},\mathbf{f}} = \frac{\sum_{e \in \mathbf{e}} \max_{f \in \mathbf{f}} s(e,f)}{|\,\mathbf{e}\,|} \quad (1)
$$

$$
\text{rec}_{\mathbf{e},\mathbf{f}} = \frac{\sum_{f \in \mathbf{f}} \max_{e \in \mathbf{e}} s(e,f)}{|\,\mathbf{f}\,|} \quad (2)
$$

$$
s_{\text{sent}} = \frac{2 \cdot \text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} \cdot \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}}}{\text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} + \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}}} \quad (3)
$$

$$
s_{i,\text{pred}} = \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \quad (4)
$$

$$
s_{i,j} = \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \quad (5)
$$

$$
\text{precision} = \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|} + w^0_{\text{nf}} s_{\text{sent}}}{\sum_i w^0_i + w^0_{\text{nf}}} \quad (6)
$$

$$
\text{recall} = \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|} + w^1_{\text{nf}} s_{\text{sent}}}{\sum_i w^1_i + w^1_{\text{nf}}} \quad (7)
$$

$$
\text{MEANT} = \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}} \quad (8)
$$

where $s(e,f)$ is the lexical similarity computed using word embeddings (Mikolov et al., 2013). By aggregating the lexical similarities, we can obtain the phrasal similarities. s_{sent} is the phrasal similarity of the whole sentence between the reference and the MT output. $s_{i,pred}$ is the phrasal similarities of the predicates between the reference translations and the MT output and $s_{i,j}$ is that of the role fillers of the arguments of role type j.

w_{pred} is the weight of the lexical similarities of the aligned predicates in step 2. w_j is the weight of the phrasal similarities of the role fillers of the arguments of role type j of the aligned frames between the reference translations and the MT output in step 3 if their role types are matching. There is a total of 12 weights for the set of semantic role labels in MEANT (Lo and Wu, 2011b) estimated by heuristics (Lo and Wu, 2012).

Finally, the weight α for the precision and recall is introduced for different usages of MEANT. α should be set to 1 so that MEANT is pure recall when it is used for MT evaluation and α should be set to 0.5 so that MEANT is the balance of precision and recall, when it is used for MT system optimization.

HMEANT (Lo and Wu, 2011a) is the variant of MEANT for human evaluation, where the semantic roles in the reference and in the MT output are annotated by humans. XMEANT (Lo et al., 2014) is the cross-lingual variant of MEANT, which estimates translation quality of the MT output against the source sentence using automatic semantic parsers for the input and output languages and alignment probabilities to determine the cross-lingual lexical semantic similarity.

3 Improvements in MEANT 2.0

We improve the performance of MEANT on evaluating translation adequacy by weighing the importance of each word by inverse document frequency when computing phrasal similarity, so that a higher score will be given to phrases with more matches for content words than for function words. We also modify the phrasal similarity calculation so that instead of aggregating lexical similarities for the bag of words in the phrase, it aggregates ngram lexical similarities. Thus, the word order of the semantic role fillers for the whole sentence is taken into account. Our development experiments showed that the optimal value of n is 2.

We also generalize the concept of weighted precision and recall when computing phrasal similarities for the semantic role fillers. Lastly, we simplify the computation of the frame semantic similarities by introducing a weight β to linearly combine the phrasal similarity of the whole sentence and the frame semantic similarity of the reference and the MT output into MEANT. Our development experiments show that the optimal value of β is 0.1. In summary, equations (1) to (8) are replaced by equations (9) to (16) as follow:

$$
\begin{aligned}
\text{prec}_{\mathbf{e},\mathbf{f}} &\equiv \text{idf-weighted max-aligned distributional ngram precision} \quad (9) \\
\text{rec}_{\mathbf{e},\mathbf{f}} &\equiv \text{idf-weighted max-aligned distributional ngram recall} \quad (10)
\end{aligned}
$$

$$
s_{i,\text{pred}} = \frac{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\alpha \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + (1-\alpha) \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \quad (11)
$$

$$
s_{i,j} = \frac{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\alpha \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + (1-\alpha) \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \quad (12)
$$

$$
s_{\text{sent}} = \frac{\text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} \cdot \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}}}{\alpha \cdot \text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} + (1-\alpha) \cdot \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} \quad (13)
$$

$$
\text{precision} = \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i} \quad (14)
$$

$$
\text{recall} = \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i} \quad (15)
$$

lang.	# sent.	# tokens	# resulted vocab.
cs	67M	1,088M	1,963k
de	184M	3,444M	4,018k
en	331M	6,585M	2,368k
fi	14M	215M	1,405k
fr	38M	1,047M	950k
hi	1M	37M	82k
lv	11M	194M	586k
pl	39M	318M	885k
ro	2M	57M	233k
ru	52M	983M	1,708k
tr	2M	34M	279k
zh	61M	2,227M	911k

Table 1: Statistics of resources used to train the word embeddings and the resulted vocabulary size of the model.

$$\text{MEANT} = \beta \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}} + (1-\beta) s_{\text{sent}} \quad (16)$$

As a result, for languages without an automatic semantic parser or sentences without a valid predicate-argument structure recognized by an automatic semantic parser, the MEANT score is the phrasal similarity of the whole sentence.

4 Setup

We use the monolingual corpora provided for the WMT translation task (Bojar et al., 2014, 2015, 2016a) to build the word embeddings for evaluating lexical similarities using `word2vec` (Mikolov et al., 2013). Table 1 summarizes the resources used to train the word embeddings and the resulting vocabulary size of the distributional lexical semantic similarity model.

We use `mateplus` (Roth and Woodsend, 2014) for German and English semantic role labeling and `mate-tools` (Björkelund et al., 2009) for Chinese semantic role labeling. Instead of the 12 semantic role types used in (Lo and Wu, 2011b), we merge the semantic role labels of Chinese, English and German into 8 role types (who, did, what, whom, when, where, why, how) for more robust performance.

For languages except Chinese, tokenization step simply involves separating punctuations at the end of the words in both the reference and the MT output. Chinese does not have clear word boundaries. Each individual Chinese character usually carries multiple meanings and relies on surrounding characters to disambiguate it. Naive Chinese character segmentation would affect the accuracy of the vector representation and the distributional lexical semantic similarity model. Thus, we use `ICTCLAS` (Zhang et al., 2003) to segment the Chinese monolingual corpus into words before building the word embeddings.

5 Experiments and results

We use the WMT 2014-2016 metrics task evaluation set (Machacek and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016b) for our development experiments. The official human judgments of translation quality were collected using relative ranking. The annotators were given the original input and the reference and were asked to order up to 5 different MT outputs according to the translation quality.

Two other kinds of human judgments of translation quality were collected in the WMT 2016 metrics task. The direct assessment evaluation protocol gave the annotators the reference and one MT output only and asked them to evaluate the translation adequacy of the MT output on an absolute scale. The HUME metric (Birch et al., 2016) is very similar to HMEANT, which evaluates translation adequacy via semantic units in the input sentence annotated by humans following the UCCA (Abend and Rappoport, 2013) guidelines. However, HUME also takes nominal and adjectival argument structures into account (instead of only predicate argument structure as in HMEANT).

Due to space limitations, we only report the results of `MEANT 2.0`, `MEANT 2.0 - nosrl`, BLEU and the best correlation in each of the individual language pairs. Since we use exactly the same protocol for each of the test sets, our reported results are directly comparable with those reported in Machacek and Bojar (2014); Stanojević et al. (2015); Bojar et al. (2016b). We summarize the observations in the following sections.

5.1 Correlation with human at system-level

5.1.1 On relative ranking judgment

Table 2 shows the Pearson's correlation with the WMT 2014-2016 official human relative ranking scores at system-level. As expected, `MEANT 2.0` performs significantly better than `MEANT 2.0 - nosrl` in most of the language pairs. Overall, both `MEANT 2.0` and the `nosrl` variant are very competitive with other metrics for all test sets.

For the WMT14 test set, `MEANT 2.0` is the best metric among all participants in that year in the de-en and en-de direction. On average over from-English directions, `MEANT 2.0`

	input language	cs	de	fr	hi	ru		en	en	en	en		en
	output language	en	en	en	en	en	ave.	cs	fr	hi	ru	ave.	de
WMT14	MEANT 2.0	.990	**.960**	.979	.791	.843	.913	–	–	–	–	–	**.482**
RRsys	MEANT 2.0 - nosrl	.983	.957	.979	.761	.830	.902	.978	.941	.986	.938	**.961**	.236
	individual best	**.993**	.943	**.981**	**.976**	**.870**	**.944**	**.988**	**.960**	**.990**	**.941**	.959	.357
	BLEU	.909	.832	.952	.956	.789	.888	.976	.937	.973	.915	.950	.216
	input language	cs	de	fi	fr	ru		en	en	en	en	en	
	output language	en	en	en	en	en	ave.	cs	de	fi	fr	ru	ave.
WMT15	MEANT 2.0	.974	.965	.946	.994	.970	.970	–	.764	–	–	–	–
RRsys	MEANT 2.0 - nosrl	.972	.956	.939	.995	.969	.966	**.984**	.676	.833	.961	.937	.878
	individual best	**.993**	**.981**	**.977**	**.997**	**.981**	**.978**	.977	**.879**	**.878**	**.964**	**.970**	**.916**
	BLEU	.957	.865	.929	.975	.851	.915	.936	.573	.602	.948	.841	.780
	input language	cs	de	fi	ro	ru	tr	en	en	en	en	en	en
	output language	en	en	en	en	en	en	cs	de	fi	ro	ru	tr
WMT16	MEANT 2.0	.989	.947	.953	.940	**.990**	.980	–	.540	–	–	–	–
RRsys	MEANT 2.0 - nosrl	.985	.928	.969	.917	.984	.978	.967	.541	.902	.868	.925	.933
	individual best	**.997**	**.985**	**.974**	**.970**	**.990**	**.981**	**.975**	**.915**	**.974**	**.959**	**.954**	**.956**
	BLEU	.992	.905	.858	.899	.962	.899	.968	.752	.868	.897	.835	.745

Table 2: Pearson's correlation of the metric scores with the WMT 2014-2016 official human relative ranking scores at system-level. For consistency with the task overview paper, en-de results are not included into out-of-English system average in WMT 2014 results (Machacek and Bojar, 2014); system average are not reported in WMT 2016 results (Bojar et al., 2016b).

	input language	cs	de	fi	ro	ru	tr	en
	output language	en	en	en	en	en	en	ru
WMT16	MEANT 2.0	.990	.950	.966	.946	.959	**.990**	–
DAsys	MEANT 2.0 - nosrl	.988	.942	.979	.930	.958	.987	.946
	individual best	**.995**	**.985**	**.980**	**.957**	**.976**	.982	**.966**
	BLEU	.989	.808	.864	.940	.837	.895	.838

Table 3: Pearson's correlation of metric scores with the WMT 2016 direct assessment of translation adequacy at system-level.

`- nosrl` is the best metric among all the participants in that year. On average over into-English directions, `MEANT 2.0` ties with the 4th-place participant in that year while `MEANT 2.0 - nosrl` is in 7th place. Both variants of MEANT lose only to ensemble and trained metrics in that year.

For the WMT15 test set, `MEANT 2.0 - nosrl` is the best metric among all the participants in that year in the en-cs direction. On average over into-English directions, `MEANT 2.0` is in 6th place while the `nosrl` variant is in 9th place. Both variants of MEANT lose only to ensemble and trained metrics in that year.

For the WMT16 test set, `MEANT 2.0` ties for 1st place with the best metric in the ru-en direction in that year. Both variants of MEANT perform as well as the leading metrics in all other directions, except en-de.

5.1.2 On direct assessment judgment

Table 3 shows the Pearson's correlation with the WMT 2016 direct assessment of translation adequacy at system-level.

Both `MEANT 2.0` and `MEANT 2.0 - nosrl` beat all the other metrics that year in the tr-en direction and perform very competitively when compared to the leading pack in other directions. `MEANT 2.0` performs better than the `nosrl` variant in all directions, except fi-en.

5.2 Correlation with human judgment at segment-level

5.2.1 On relative ranking judgment

Table 4 shows the Kendall's correlation with the WMT 2014-2016 official human relative ranking judgments at segment-level. Similar to the correlation at the system-level, `MEANT 2.0` performs significantly better than `MEANT 2.0 - nosrl` for most language pairs.

For the WMT14 test set, `MEANT 2.0` beats all the participants in the en-de direction while the `nosrl` variant beats all the participants in all other from-English directions (and their average) in that year. On the average of all the to-English directions, `MEANT 2.0` and the `nosrl` variant are in 2nd and 3rd place respectively and only lose to an ensemble and trained metric in that year.

	input language	cs	de	fr	hi	ru		en	en	en	en	en	
	output language	en	en	en	en	en	ave.	cs	de	fr	hi	ru	ave.
WMT14 RRseg	MEANT 2.0	.325	.353	.421	.421	.348	.374	–	**.279**	–	–	–	–
	MEANT 2.0 - nosrl	.312	.354	.426	.410	.341	.367	**.355**	.254	**.314**	**.294**	**.472**	**.338**
	individual best	**.328**	**.380**	**.433**	**.438**	**.355**	**.386**	.344	.268	.293	.286	.440	.319
	sentBLEU	.213	.271	.378	.300	.263	.285	.290	.191	.256	.227	.381	.269
	input language	cs	de	fi	fr	ru		en	en	en	en	en	
	output language	en	en	en	en	en	ave.	cs	de	fi	fr	ru	ave.
WMT15 RRseg	MEANT 2.0	.463	.465	.424	.402	.400	.431	–	.398	–	–	–	–
	MEANT 2.0 - nosrl	.463	.454	.421	**.406**	.401	.429	**.472**	.386	.344	.365	**.442**	**.402**
	individual best	**.495**	**.482**	**.445**	.398	**.418**	**.447**	.446	**.399**	**.380**	**.366**	.439	.400
	sentBLEU	.391	.360	.308	.358	.329	.349	.290	.191	.256	.227	.381	.269
	input language	cs	de	fi	ro	ru	tr	en	en	en	en	en	en
	output language	en	en	en	en	en	en	cs	de	fi	ro	ru	tr
WMT16 RRseg	MEANT 2.0	.355	**.453**	.414	.345	.401	.373	–	**.360**	–	–	–	–
	MEANT 2.0 - nosrl	.347	.438	.411	.338	.400	.364	**.436**	**.360**	.329	.271	**.428**	.325
	individual best	**.388**	.420	**.481**	**.383**	**.420**	**.424**	.422	.334	**.364**	**.307**	.405	**.337**
	sentBLEU	.284	.265	.368	.272	.330	.245	.359	.236	.306	.233	.328	.222

Table 4: Kendall's correlation of metric scores with the WMT 2014-2016 official human relative ranking judgments at segment-level. For consistency with the task overview paper, system averages are not reported in WMT 2016 results (Bojar et al., 2016b).

	input language	cs	de	fi	ro	ru	tr	en
	output language	en	en	en	en	en	en	ru
WMT16 DAseg	MEANT 2.0	.674	.510	.539	.607	.535	.588	–
	MEANT 2.0 - nosrl	.672	.484	.522	.587	.540	.577	.664
	individual best	**.713**	**.601**	**.598**	**.661**	**.618**	**.663**	**.666**
	sentBLEU	.557	.448	.484	.499	.502	.532	.550

Table 5: Pearson's correlation of metric scores with the WMT 2016 direct assessment of absolute translation adequacy at segment-level.

For the WMT15 test set, both `MEANT 2.0` and `MEANT 2.0 - nosrl` beat all the participating metrics in that year in the fr-en direction. `MEANT 2.0 - nosrl` also has the highest correlation with human in the en-cs and en-ru directions and the overall average of the from-English directions. Again, on the average of all the to-English directions, `MEANT 2.0` and the `nosrl` variant are in 2nd and 3rd place respectively and only lose to an ensemble and trained metric.

For the WMT16 test set, both `MEANT 2.0` and `MEANT 2.0 - nosrl` beat all other participants in that year in the de-en, en-de directions while `MEANT 2.0 - nosrl` is also the champion in the en-cs and the en-ru directions. Both variants perform very competitively when compared to the leading metrics in all other directions.

5.2.2 On direct assessment judgment

Table 5 shows the Pearson's correlation of MEANT with the WMT 2016 direct assessment of absolute translation adequacy at segment-level. Both variants of MEANT perform very competitively when compared to the leading pack in other directions. `MEANT 2.0` performs better than the `nosrl` variant in all directions, except ru-en.

	input language	en	en	en	en
	output language	cs	de	pl	ro
HUME	MEANT 2.0	–	**.522**	–	–
	MEANT 2.0 - nosrl	.508	**.522**	.619	**.479**
	individual best	**.544**	.480	**.639**	.435
	sentBLEU	.349	.377	.550	.328

Table 6: Pearson's correlation of metric scores with the WMT 2016 HUME human assessment at segment-level.

5.2.3 On HUME evaluation

Table 6 shows the Pearson's correlation of MEANT with the HUME human assessment on the `himltest` test set at segment-level.

Both `MEANT 2.0` and `MEANT 2.0 - nosrl` beat all other participating metrics in that year in the en-de direction. `MEANT 2.0 - nosrl` also has the highest correlation with HUME among all the participants in that year in the en-pl direction.

5.3 Evaluation speed

Table 7 shows the average time (in seconds) for each step of a typical WMT system evaluation

lang.	# pairs	tok.	load	srl	score
cs	2.8k	3	158	–	56
de	2.6k	2	333	1010	41
en	2.8k	2	195	1120	46
fi	2.3k	2	114	–	24
fr	3.0k	4	77	–	61
hi	2.5k	4	7	–	35
lv	2.0k	2	47	–	13
pl	0.3k	1	72	–	4
ro	2.0k	1	19	–	10
ru	2.9k	5	142	–	27
tr	3.0k	4	23	–	17
zh	2.0k	501	75	1175	16

Table 7: Average time in seconds for each step of evaluating a typical WMT system using MEANT: tokenizing both the reference and the MT output; loading the distributional lexical semantic similarity model; semantic role labeling the reference and the MT output; and scoring the MT output.

on different output languages using MEANT. The time taken for punctuation tokenization is almost negligible. This is because in common practice for MT system development, the validation and evaluation set are reused frequently, so the processing of the reference translation is typically precomputed. Furthermore, the MT system is trained to output tokenized translations, so it is not necessary to run the tokenization step on the MT output. Therefore, the tokenization step does not affect the time cost of MEANT in practical applications (even in the case of Chinese, where word segmentation takes significantly longer).

The loading time of the word embedding model is proportional to the vocabulary size of the model reported in table 1; it takes less than a second to load 10k vocabularies into memory.

Automatic semantic role labeling (SRL) is the most time consuming step in running MEANT. The time reported in table 7 includes parsing both the reference and the MT output. However, as pointed out above, common practice for MT system development is to frequently reuse the validation and evaluation sets. Thus, semantic role labeling of the reference translation could be precomputed to reduce the time taken for the SRL step in the development cycle.

Finally, the time used in computing the MEANT score is proportional to the size of the evaluation set and the word embedding model. The scoring step processes around 50 to 100 sentences each second.

6 Conclusion

We present a new version of MEANT that participated in evaluating MT systems for all language pairs in the metrics task of the Second Conference on Machine Translation (WMT 2017). The correlation of MEANT with human judgment has been improved by better addressing translation adequacy via weighing the importance of each word in the phrasal similarity computation by inverse document frequency, and better addressing translation fluency via using distributional ngram accuracy to account for word reordering in the computation. Our results show that MEANT consistently performs well across different output languages in the previous year's test set at both system-level and segment-level.

MEANT 2.0 - nosrl is a non-ensemble and untrained metric that requires only a monolingual corpus in the output language for building the word embeddings to evaluate MT systems for a new language pair. Although there is an overhead time cost in semantic role labeling sentence pairs in MEANT 2.0 and loading the word embedding model in both MEANT 2.0 and its nosrl subversion, the time cost can be reduced almost by half in real applications. This highly portable and open source semantic MT evaluation metric is a more accurate alternative to BLEU in evaluating translation quality for low-resource languages.

Acknowledgement

The author would like to thank Markus Saers, Karteek Addanki and Meriem Beloucif for providing code review before the software release and Roland Kuhn for editing the paper.

References

Omri Abend and Ari Rappoport. 2013. UCCA: A Semantics-based Grammatical Annotation Scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Association for Computational Linguistics, Potsdam, Germany, pages 1–12. http://www.aclweb.org/anthology/W13-0101.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-Based Evaluation of Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1264–1274. https://aclweb.org/anthology/D16-1134.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Association for Computational Linguistics, Boulder, Colorado, pages 43–48. http://www.aclweb.org/anthology/W09-1206.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 12–58. http://www.aclweb.org/anthology/W14-3302.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W16-2301.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46. http://aclweb.org/anthology/W15-3001.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. 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.

Michael Denkowski and Alon Lavie. 2014. METEOR universal: Language specific translation evaluation for any target language. In *9th Workshop on Statistical Machine Translation (WMT 2014)*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*. San Diego, California.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDer: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 765–771. https://doi.org/10.3115/v1/P14-2124.

Chi-kiu Lo, Philipp Dowling, and Dekai Wu. 2015. Improving evaluation and optimization of MT systems against MEANT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 434–441. http://aclweb.org/anthology/W15-3056.

Chi-kiu Lo and Dekai Wu. 2011a. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 220–229. http://aclweb.org/anthology/P11-1023.

Chi-Kiu Lo and Dekai Wu. 2011b. SMT Versus AI Redux: How Semantic Fames Evaluate MT More Accurately. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*. AAAI Press, IJCAI'11, pages 1838–1845. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-308.

Chi-kiu Lo and Dekai Wu. 2012. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Jeju, Republic of Korea, pages 49–56. http://www.aclweb.org/anthology/W12-4206.

Matous Machacek and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 293–301. http://www.aclweb.org/anthology/W14-3336.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, NIPS'13, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method

for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, Pennsylvania, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 407–413. http://www.aclweb.org/anthology/D14-1045.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*. Cambridge, Massachusetts, pages 223–231.

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.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics, Sapporo, Japan, pages 184–187.