Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation

Anand Karthik TUMULURU, Chi-kiu LO and **Dekai WU**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo,aktumuluru,dekai}@cs.ust.hk

Abstract

We present larger-scale evidence overturning previous results, showing that among the many alternative phrasal lexical similarity measures based on word vectors, the Jaccard coefficient most increases the robustness of MEANT, the recently introduced, fully-automatic, state-of-the-art semantic MT evaluation metric. MEANT critically depends on phrasal lexical similarity scores in order to automatically determine which semantic role fillers should be aligned between reference and machine translations. The robustness experiments were conducted across various data sets following NIST MetricsMaTr protocols, showing higher Kendall correlation with human adequacy judgments against BLEU, METEOR (with and without synsets), WER, PER, TER and CDER. The Jaccard coefficient is shown to be more discriminative and robust than cosine similarity, the Min/Max metric with mutual information, Jensen Shannon divergence, or the Dice's coefficient. We also show that with Jaccard coefficient as the phrasal lexical similarity metric, individual word token scores are best aggregated into phrasal segment similarity scores using the geometric mean, rather than either the arithmetic mean or competitive linking style word alignments. Furthermore, we show empirically that a context window size of 5 captures the optimal amount of information for training the word vectors. The combined results suggest a new formulation of MEANT with significantly improved robustness across data sets.

1 Introduction

We present larger-scale evidence overturning previous results, showing that the Jaccard coefficient among the alternative lexical similarity measure based on word vectors most increases the robustness of MEANT, even more than that of the Min/Max metric with mutual information metric, as used by Lo *et al.* (2012) in their formulation of MEANT that outperformed BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann *et al.*, 1997), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006).

MEANT, the fully-automatic, state-of-the-art semantic MT evaluation metric as introduced by Lo *et al.* (2012) uses the Min/Max metric with mutual information on word vectors as the similarity measure to score phrasal similarity of the semantic role fillers which is the matching criterion to align semantic frames. In achieving the same, word vectors are trained on a window size of 5 and use arithmetic mean to aggregate token similarity scores into segment similarity scores.

We explore the potential of alternate similarity metrics on word vectors such as the Jensen Shannon divergence, the Dice's coefficient and Jaccard coefficient apart from cosine similarity and the Min/Max metric with mutual information employed by Lo *et al.* (2012) in their work. We show that Jaccard coefficient not only outperforms the Min/Max metric with mutual information, in achieving higher Kendall correlation against human adequacy judgments, but all the other similarity measures in comparison.

In order to test the robustness of the method across various data sets, we conduct experiments across GALE-A, GALE-B and GALE-C data sets examining the Kendall correlation against human adequacy judgments following NIST MetricsMaTr protocols (Callison-Burch *et al.*, 2010). We train the weights used for computing the weighted f-score over matching role labels using a grid search and then test them on a combination of these data sets and since each data set has different average sentence length and number of sentences we identify robust metrics that perform across all the variations after thorough analysis on the quality of the weights assigned to the role labels.

The strategy used in evaluating the phrasal similarity score from the component token similarity scores is critical in deciding the overall performance of the MEANT metric, as role fillers are often phrases. In contrast to the arithmetic mean and competitive linking strategies we show that that using the geometric mean for this purpose

26th Pacific Asia Conference on Language,Information and Computation pages 574–581

is more reliable.

In order to examine the optimum amount of contextual information to be captured while training the word vectors, we vary the window size while training the word vectors from 3 to 13. Surprisingly, we achieve both high performance and robustness at the window size of 5 not only for Jaccard coefficient but across almost all the metrics in comparison.

Our results indicate that Jaccard coefficient on word vectors trained with a window size of 5, and using geometric mean style of aggregation as the criterion for aligning semantic frames and significantly enhances the performance in comparison to other metrics and robustness across varying data sets of MEANT.

2 Related work

Evaluating lexical similarity of phrases plays an important role in many language technology applications such as Machine Translation Evaluation, Word Sense disambiguation, Query Expansion, Information Retrieval, Question Answering etc.

BLEU (Papineni *et al.*, 2002) , NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann *et al.*, 1997), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006), are some of the commonly used phrasal similarity metrics. Although lexical similarity evaluation with all the metrics can be done very quickly at low cost, they assume that a good translation shares the same lexical choices as the reference translation, which is not justified semantically.

We argue that a lexical similarity metric that reflects meaning similarity needs to be aware of the contextual similarity, and not merely flat lexical similarity.

3 Word vector models and similarity metrics

Word Vector models (Dagan, 2000) are guided by the principle that similar words occur in similar contexts. In the word vector model, each word in the lexicon is represented by a word vector, where each entry corresponds to the frequency of cooccurence with every other word in the lexicon. The definition of the cooccurence relation decides the nature of the context we capture and have been used in a wide variety of tasks, such as in word sense disambiguation by Gale *et al.* (1992) by defining the relation as the cooccurence within a distance of 50 words. Grammatical and syntactic relations were also identified, by defining the relation as the cooccurence in a relatively shorter window of 5 words, as in the work of Smadja (1993) and Dagan *et al.* (1993) . The word vector models can be readily trained on any large mono-lingual corpora and hence their utility is not constrained to resource rich languages.

In this work, we make a choice of defining the cooccurence relation as the joint cooccurence of the word within a short window of text, by the principle of occam's razor. A window size of n symmetrically encompasses word tokens at a distance of upto $\frac{(n-1)}{2}$ on both directions and hence captures not only semantic context, but but also a mixture of grammatical and topical cooccurences. We make a choice of not using any techniques such as stemming, lemmatisation or stop-word pruning as using such limit the use of the word vector models to only some languages.

The trained word vectors can be used with a variety of mathematical measures of similarity between a pair of vectors to evaluate the degree of similarity of the words that they represent. We use a diverse set of such functions, each quantifying a different aspect of the accumulated cooccurence statistics between a pair of vectors.

3.1 Cosine Similarity

Cosine measure gives the cosine of the angle between the two vectors and is commonly used in the vector space model. Since the word vectors have non-negative components, the range is between 0 and 1, where a value of 0 indicates that the vectors are orthogonal or dissimilary and a value of 1 indicates that the vectors are parallel or similar. The cosine similarity between two tokens x and y is defined as follows:

$$\begin{aligned} \overrightarrow{w_x} &= \text{context vector of word token x} \\ w_{xi} &= \text{attribute i of context vector } \overrightarrow{w_x} \end{aligned}$$

$$f(x, w_{xi}) = \frac{c(x, w_{xi})}{\sum_j c(x, w_{xj})}$$

$$\text{cosine}(x, y) = \frac{\sum_i f(x, w_{xi}) \times f(y, w_{yi})}{\sqrt{\sum_i f(x, w_{xi})^2} \sqrt{\sum_i f(y, w_{yi})^2}}$$

3.2 Min/Max metric with Mutual Information

Using the above given definition of w_{xi}, the min/max with mutual information (Cover and Thomas, 1991) similarity between two sequences of two tokens, x and y is defined as follows:

$$P(w_{xi} \mid x) = \frac{c(x, w_{xi})}{\sum_j c(x, w_{xj})}$$

$$P(w_{xi}) = \frac{\sum_y c(y, w_{xi})}{\sum_y \sum_j c(y, w_{xj})}$$

$$\text{MI}(x, w_{xi}) = \log\left(\frac{P(w_{xi} \mid x)}{P(w_{xi})}\right)$$

$$\text{MinMax-MI}(x,y) = \frac{\sum_i \min\left(\text{MI}(x, w_{xi}), \text{MI}(y, w_{y_i})\right)}{\sum_i \max\left(\text{MI}(x, w_{xi}), \text{MI}(y, w_{y_i})\right)}$$

The range of Min/Max metric with Mutual Information is 0 to 1, a value of 0 indicates that the vectors are completely dissimilar and a value of 1 indicated that they are identical.

3.3 Jensen Shannon Divergence

Using the above given definitions of w_{xi}, the Jesen Shannon divergence (Lin, 1991), (Rao, 1982) is defined as follows:

$$D(x \mid\mid \frac{x+y}{2}) = \sum_i P(w_{x_i} \mid x) \log\left(\frac{2 \times P(w_{x_i} \mid x)}{P(w_{x_i} \mid x) + P(w_{y_i} \mid y)}\right)$$

$$\text{JSD}(x,y) = D(x \mid\mid \frac{x+y}{2}) + D(y \mid\mid \frac{x+y}{2})$$

Here, $D(x \mid\mid y)$ represents the Kullback-Leibler Divergence (Cover and Thomas, 1991). The Jensen Shannon divergence addresses the problem of assymetry assosiated with KL divergence, and has a range of 0 to 1. The square root of Jensen Shannon Divergence is a metric, also with a range of 0 to 1, but since it is divergence metric, a value of 0 indicates that the vectors of x and y are similar and a value of 1 indicates that they are orthogonal.

3.4 Dice's coefficient

Dice's coefficient for two words x and y is defined as the ratio of total number of shared cooccurences of their vectors to the total number of cooccurences in both the vectors. It is formulated as follows:

$$\text{DC}(x,y) = \frac{\sum_i \min\left(c(x, w_{xi}), c(y, w_{y_i})\right)}{\sum_i \left(c(x, w_{xi}) + c(y, w_{y_i})\right)}$$

where the definitions of w_{xi} and $c(x, w_{x_i})$are the same as above. Here, $\min(a, b)$represents the minimum of the values a, b. The range of Dice's coefficient is 0 to 1, a value of 0 indicates that the vectors are completely dissimilar and a value of 1 indicated that they are identical.

3.5 Jaccard coefficient

The Jaccard coefficient for two words x and y is defined as the ratio of intersection of their cooccurences to the union of their cooccurences of their word vectors.

$$\text{JC}(x,y) = \frac{\sum_i \min\left(c(x, w_{xi}), c(y, w_{y_i})\right)}{\sum_i \max\left(c(x, w_{xi}), c(y, w_{y_i})\right)}$$

where the definitions of w_{x_i}, $c(x, w_{x_i})$ and $\min(a, b)$ are the same as above. Here, $\max(a, b)$ represents the maximum of the values a, b

The range of Jaccard coefficient is 0 to 1, a value of 0 indicates that the vectors are completely dissimilar and a value of 1 indicated that they are identical.

4 Computing phrasal similarity

In this section, we define the methods used in computing the similarity of two phrases given the degree of similarity of the component tokens. Evaluating phrasal similarity in the context of word vectors is a challenge, as we have no information about the alignment of the token pairs in the given phrases.The strategy employed must provide sufficient discriminatory power in order for MEANT to align the one pair of similar role fillers among many such pairs with mismatched lengths and word ordering. We now discuss the methods we use in computing the phrasal similarity scores from the component token similarity scores.

4.1 Arithmetic Mean

In this method, we simply assume that there is a complete alignment between the two phrases. We then compute the phrasal similarity score as the mean of similarity scores of all the component token pairs. The phrasal similarity between two sequences of word tokens $\overrightarrow{u}$ and $\overrightarrow{v}$ using the arithmetic mean method is defined as:

$$\text{AM}(\overrightarrow{u}, \overrightarrow{v}) = \frac{1}{t \times s} \sum_i \sum_j S(u_i, v_j)$$

where t is the number of word tokens in $\overrightarrow{u}$ and s is the number of word tokens in $\overrightarrow{v}$. $S(u_i, v_j)$ is the token similarity score of the i^{th} token in $\overrightarrow{u}$ and the j^{th} token in $\overrightarrow{v}$ obtained using any of the above mentioned token similarity metrics.

4.2 Geometric Mean

In this method, again, we assume that there a complete alignment between the two phrases. We then compute the phrasal similarity score as the geometric mean of similarity scores of all the component token pairs. The phrasal similarity between two sequences of word tokens $\overrightarrow{u}$ and $\overrightarrow{v}$ using the geometric mean method is defined as:

$$\text{GM}(\overrightarrow{u}, \overrightarrow{v}) = e^{\frac{1}{(t \times s)} \sum_j \sum_i \ln(S(u_i, v_j))}$$

where t is the number of word tokens in $\overrightarrow{u}$ and s is the number of word tokens in $\overrightarrow{v}$. $S(u_i, v_j)$is the token similarity score of the i^{th}token in $\overrightarrow{u}$and the j^{th} token in $\overrightarrow{v}$ obtained using any of the above mentioned token similarity metrics.

4.3 Modified Competitive Linking

In this method we attempt to align the tokens in the phrases using the similarity score of the token pair as a heuristic. As in the previous methods, we avoid the danger of aligning a token in one segment to excessive numbers of tokens in the other segment, by adopting a variant of competitive linking by Melamed (1996). The competitive linking algorithm adopts a greedy best first strategy

in making strictly one to one word alignments. Since we frequently encounter phrases for alignment with unequal lengths, this one to one constraint severely restricts alignments and so we modify the competitive linking strategy by allowing one to many alignments. The number of such one to many alignments must be equal to the difference in the segment lengths. Once these alignments have been made, we compute the similarity of the two phrases as the arithmetic mean of the similarity scores of the aligned tokens.

5 Jaccard coefficient outperforms other metrics

We show that the Jaccard coefficient outperforms other similarity metrics as the criterion for evaluating lexical similarity to align role fillers in MEANT.

5.1 Experimental Setup

We report the performance of all the similarity metrics - cosine similarity, Min/Max with mutual information, Jensen Shannon divergence, Jaccard coefficient and the Dice's coefficient on the word vector models as described above as criterion for aligning semantic frames in MEANT.

We train the word vector models on the uncased Gigaword corpus. We do not use techniqes such as stemming, lemmatisation or stop-word pruning. We train the word vectors on the Gigaword corpus with window sizes ranging from 3 to 13.

For our benchmark comparison, the evaluation data for our experiments is the same two sets of sentences, GALE-A and GALE-B that were used in Lo and Wu (2011), where in GALE-A is used for estimating the weight parameters of the metric by optimizing the correlation with human adequacy judgment, and then the learned weights are applied to testing on GALE-B. For the automatic semantic role labeling, we used the publicly available off-the-shelf shallow semantic parser, ASSERT (Pradhan *et al.*, 2004). Semantic frame alignment is done by applying maximum bipartite matching algorithm with the lexical similrity of predicates as edge weights. The correlation with human adequacy judgments on sentence-level system ranking is assessed by the standard NIST MetricsMaTr procedure (Callison-Burch *et al.*, 2010) using Kendall correlation coefficients.

We first run a grid search on the GALE-A data set for each of these metrics on all window sizes to obtain weights for the role labels. We then use these weights to evaluate the GALE-C data set. The Kendall correlation score is obtained using MEANT as described in Lo Wu 2012. In this experiment, we use geometric mean as the aggregation method and vary the window sizes for each metric to first identify one metric that performs robustly across all window sizes for the given dataset. We also examine the distribution of weights over the semantic role labels across all the window sizes to verify that the metric is both: performing consistently and producing the excpected distribution of weights over semantic role labels.

5.2 Results

Table 1 shows that the Jaccard coefficient performs consistently well and relatively out performs most other similarity metrics in comparision. It is surprising that the performance of all the metrics does not improve significantly and sometimes, decreases with increasing window size. For a window size of 5 for the Jaccard coefficient, we achieve close to 0.21 Kendall for testing on GALE-B, outperforming the scores reported on the same data sets MEANT in Lo *et al.* (2012). A kendall of 0.26 and 0.22 are observed for Dice's coefficient with window sizes of 3 and 11. On a closer look at the weights assigned to each role labels after training on GALE A, we observe that the weights in these cases have been abnormally chosen in the favour of matching role fillers with less important role labels, but on the contarary, in the case of Jaccard coefficient they have been distributed with relatively higher importance for predicate, agr0, arg1 and arg2 across all window sizes indicating that it is enabling the alignment of more important roles accurately.

6 Phrasal similarity best computed through geometric mean

We show that the Geometric mean method of aggregation out performs the arithmetic and competitive linking methods using Jaccard coefficient

6.1 Experimental Setup

We report the performance of the arithmetic mean and competitive linking methods of aggregation using Jaccard coefficient as the lexical similarity measure. These similarity metrics are employed on word vectors trained on the Gigaword corpus with window sizes ranging from 3 to 13. The evaluation data for our experiment is the same as described above.

6.2 Results

In tables 2 and 3, we observe that the geometric mean method of aggregation outperforms arithmetic mean and competitive linking methods of aggregation. Although we see markedly higher kendall scores with training on the GALE-A data set using the modified competitive linking method of aggregation, the resultant weights that yield such high scores are not only improperly distributed, but also perform poorly when tested on the

Table 1: Kendall correlation scores with human adequacy judgment on GALE-A (training) and GALE-B (testing) comparing MEANT integrated with various lexical similarity measures as criterion for aligning semantic role fillers: (a) cosine similarity, (b) Min/Max with mutual information (c) Jensen Shannon divergence (d) Jaccard coefficient and (e) Dice's coefficient with word vectors trained from window sizes 3-13 and using geometric mean as the aggregation method

	Training on GALE A	Testing on GALE B
window size 3	0.2702	0.2095
window size 5	0.3783	0.1523
window size 7	0.3783	0.1142
window size 9	0.3153	0.0857
window size 11	0.2972	0.180
window size 13	0.3603	0.1523
Min/Max with MI		
window size 3	0.3603	0.1333
window size 5	0.3603	0.1523
window size 7	0.2252	0.1714
window size 9	0.3333	0.2476
window size 11	0.2882	0.1523
window size 13	0.2522	0.1142
JSD		
window size 3	0.3963	0
window size 5	0.3603	0
window size 7	0.3423	0
window size 9	0.3603	0
window size 11	0.3243	0.0952
window size 13	0.3603	0.1428
Jaccard Coefficient		
window size 3	0.3783	0.1904
window size 5	0.3333	0.2095
window size 7	0.3423	0.2000
window size 9	0.3423	0.1809
window size 11	0.3513	0.0952
window size 13	0.3513	0.1142
Dice's Coefficient		
window size 3	0.3603	0.2666
window size 5	0.3603	0.1809
window size 7	0.3513	0.1904
window size 9	0.3693	0.1714
window size 11	0.3693	0.2285
window size 13	0.3603	0.1714

Table 2: Sentence-level correlation with human adequacy judgment on GALE-A (training) and GALE-B (testing) comparing MEANT integrated with Jaccard coefficient as measure of lexical similarity on word vectors trained on window sizes 3-13 between semantic role fillers using arithmetic mean as the aggregation method

	Training on GALE A	Testing on GALE B
Jaccard Coefficient		
window size 3	0.3603	0.1523
window size 5	0.3333	0.2000
window size 7	0.3603	0.1809
window size 9	0.3603	0.1619
window size 11	0.3603	0.2380
window size 13	0.3603	0.2095

Table 3: Kendall correlation scores with human adequacy judgment on GALE-A (training) and GALE-B (testing) comparing MEANT integrated with Jaccard coefficient as measure of lexical similarity on word vectors trained with window sizes 3-13 between semantic role fillers using Competitive linking as the aggregation method

	Training on GALE A	Testing on GALE B
Jaccard coefficient		
window size 3	0.3873	0.1714
window size 5	0.3963	0.1904
window size 7	0.3783	0.1333
window size 9	0.3783	0.0952
window size 11	0.3693	0.0761
window size 13	0.3693	0.0952

Table 4: Kendall correlation scores with human adequacy judgment and the corresponding role label weights on GALE-C as the training set and GALE-A as the testing set with MEANT integrated with Jaccard coefficient as measure of lexical similarity between semantic role fillers with word vectors trained on window sizes 3-13 and using Geometric Mean as the aggregation method. The role labels are pr - predicate, a0 - arg0, a1 - arg1, a2 - arg2, te - temporal, lo - locative, pu - purpose, ex - extent, ma - manner, o - other, m - model, n - negation

Jaccard Coefficient	GALE C	GALE A	pr	a0	a1	a2	te	lo	pu	ex	ma	o	m	n
window size 3	0.1443	0.1981	2	3	1	0	2	0	2	0	0	0	0	2
window size 5	0.1520	0.3243	5	3	0	0	1	0	0	1	0	1	0	1
window size 7	0.1505	0.1351	0	4	4	0	0	0	3	0	0	0	0	1
window size 9	0.1520	0.1441	1	2	0	0	3	0	1	0	0	2	0	3
window size 11	0.1505	0.1441	1	2	0	0	3	0	1	0	0	2	0	3
window size 13	0.1566	0.1441	1	2	0	0	3	0	1	0	0	2	0	3

GALE-B data set. Other variants of the competitive linking method of similar nature may also be expected to suffer from this problem of overfitting.

The arithmetic mean method performs extremely well in the case of higher window sizes - 11 and 13 in this pirticular case, where we use GALE-A as the train data set and GALE-B as the test dataset, but does not perform as well as the geometric mean over relatively larger datasets as in the case of training with GALE-C and testing on GALE-A and GALE-B, where we observe negative correlation scores.

It has been observed, the method in which we compute the phrasal similarity scores from the component token similarity scores of the role fillers impacts the overall performance at two levels - (1) In effectively handling different lengths of phrases and (2) In the distribution of weights on the roles. By out-performing arithmetic mean and competitive linking, the geometric mean method of aggregation as seen in table 1 has proven to handle both the factors robustly.

7 Jaccard coefficient is robust across various data sets

Given the positive results on the above mentioned data sets, we ask : Does Jaccard coefficient perform robustly across various data sets? The concerns with varying data sets is two fold: (1) Does Jaccard coefficient as a metric have enough discriminatory power? (2) Is the Jaccard coefficient enabling consistent distribution of weights to role labels during training.

7.1 Experimental Setup

We follow a similar setup as laid out in the previous experiments, except for our benchmark comparison, the evaluation data for our experiments we use GALE-A, GALE-B and GALE-C as used in that were used in Lo and Wu (2011), where in GALE-C is used for estimating the weight parameters of the metric by optimizing the correlation with human adequacy judgment, and then the learned weights are applied to testing on both GALE-A and GALE-B.

7.2 Results

In tables 4 and 5, we observe that Jaccard coefficient still performs very well on varying the training and testing data sets, achieving scores of 0.15, 0.32 and 0.26 on GALE-C (training), GALE-A (testing) and GALE-B (testing) respectively. This indicates robustness of Jaccard coefficient as a lexical metric and its reliability for using it across any new data sets.

Table 5: Kendall correlation scores with human adequacy judgment and the corresponding role label weights on GALE-C as the training set and GALE-B as the testing set with MEANT integrated with Jaccard coefficient as measure of lexical similarity between semantic role fillers with word vectors trained on window sizes 3-13 and using Geometric Mean as the aggregation method. The role labels are pr - predicate, a0 - arg0, a1 - arg1, a2 - arg2, te - temporal, lo - locative, pu - purpose, ex - extent, ma - manner, o - other, m - model, n - negation

Jaccard Coefficeint	GALE C	GALE B	pr	a0	a1	a2	te	lo	pu	ex	ma	o	m	n
window size 3	0.1443	0.1333	2	3	1	0	2	0	2	0	0	0	0	1
window size 5	0.1520	0.2666	5	3	0	0	1	0	0	1	0	1	0	1
window size 7	0.1505	0.0476	0	4	4	0	0	0	3	0	0	0	0	1
window size 9	0.1520	0.1904	1	2	0	0	3	0	0	0	0	2	0	4
window size 11	0.1505	0.1523	1	2	0	0	3	0	1	0	0	2	0	3
window size 13	0.1566	0.1714	1	2	0	0	3	0	0	0	0	2	0	4

7.3 What is the optimal window size?

A closer analysis at the weights assigned to the role labels on training with Jaccard coefficient across all window sizes using the geometric mean method of aggregation shows that evaluating with word vectors trained on a window size of 5 gives relatively higher importance by by concentrating the weight mass over the more important role labels. This has been observed even for the experiments with a different data set - by training on GALE-A and testing on GALE-B. We also observe relatively higher scores consistent with the weighing scheme, using this combination for all the data sets. Jaccard coefficient with word vectors trained on a window size of 5, using the geometric mean method of evaluating phrasal similarity out performs all the other methods and robustly so across various data sets.

8 Conclusion

We have shown through a broad range of comparative experiments that Jaccard coefficient as a phrasal lexical similarity metric within MEANT out performs all the other metrics and most importantly than that of the Min/Max with mutual information metric, as used by Lo *et al.* (2012) in thier formulation of MEANT that outperformed BLEU, METEOR, WER, PER, CDER and TER.

We have also shown that using a window size of 5 the word vectors is optimal to train the word vectors after analyzing the performance of Jaccard coefficent across window sizes 3 to 13. Jaccard coefficient is shown to be more discriminative using the geometric mean method of aggregation over the arithmetic mean and competitive linking methods. Through experiments across across various data sets Jaccard coeffficient as a lexical similarity metric is shown to be robust and consistently yielding high performance.

By incorporating Jaccard coefficient as the lexical similarity metric, we except that the new formulation of the MEANT metric would show improved performance and robustness.

Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF621008, GRF612806, DAG03/04.EG09, RGC6256/00E, and RGC6083/99E. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the RGC, EU, or DARPA.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 65–72, 2005.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Pryzbocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, 15-16 July 2010.

T. Cover and J. Thomas. *Elements of information theory*. Wiley, New York, 1991.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. Contextual word similarity and estimation from sparse data. In Lenhart K. Schubert, editor, *ACL*, pages 164–171. ACL, 1993.

Ido Dagan. Contextual word similarity. In Robert Dale, Herman Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 459–476. Marcel Dekker, New York, 2000.

G. Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT-02)*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

William A. Gale, Kenneth Ward Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 2006.

J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145 –151, jan 1991.

Chi-kiu Lo and Dekai Wu. Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in Statistical Translation (SSST-5)*, 2011.

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

I. Dan Melamed. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, 1996.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, 2000.

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

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, 2004.

C. Radhakrishna Rao. Diversity: Its Measurement, Decomposition, Apportionment and Analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 44(1):1–22, 1982.

F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, 2006.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97)*, 1997.