Semantic Evaluation of Machine Translation

Billy Tak-Ming Wong
Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
E-mail: ctbwong@cityu.edu.hk

Abstract

It is recognized that many evaluation metrics of machine translation in use that focus on surface word level suffer from their lack of tolerance of linguistic variance, and the incorporation of linguistic features can improve their performance. To this end, WordNet is therefore widely utilized by recent evaluation metrics as a thesaurus for identifying synonym pairs. On this basis, word pairs in similar meaning, however, are still neglected. We investigate the significance of this particular word group to the performance of evaluation metrics. In our experiments we integrate eight different measures of lexical semantic similarity into an evaluation metric based on standard measures of unigram precision, recall and F-measure. It is found that a knowledge-based measure proposed by Wu and Palmer and a corpus-based measure, namely Latent Semantic Analysis, lead to an observable gain in correlation with human judgments of translation quality, in an extent to which better than the use of WordNet for synonyms.

1. Introduction

Since the proposal of BLEU (Papineni et al., 2001) and subsequent metrics, a paradigm shift occurred in the evaluation of machine translation (MT) which turns it into an automatic task from manual work. In return, automatic evaluation metrics serve as a standard benchmark for MT system performance. An improvement in metric score is conceived as an indicator of better quality of MT outputs.

In recent years, however, the reliability of the evaluation metrics has been questioned. In some cases these metrics fail to provide an appropriate assessment of MT performance. Callison-Burch et al. (2006;2007) present substantial examples that BLEU tend to underestimate the translation quality of rule-base systems. Besides, Babych and Hartley (2008) demonstrate that BLEU loses sensitivity on higher quality MT outputs. Such findings reveal the bottleneck of current MT evaluation practices relying on metrics that merely measure lexical identity at surface text level, and are insensitive to variation in further linguistic levels. Although the use of multiple references can alleviate this problem by providing different versions of translation in equivalent meaning, it is unlikely that all of the possible translations can be completely enumerated.

Some recent metrics try to deal with this problem by lessening the sole reliance on exact word match. Different kinds of linguistic analysis are incorporated into the metrics in order to account for the variance between MT outputs and human references in syntactic or semantic level. Within those, a light semantic resource, WordNet, is widely adopted by different metrics as a thesaurus to allow matching of synonyms, for instances, METEOR (Banerjee & Lavie, 2005), MAXSIM (Chan & Ng, 2008), TERp (Snover et al., 2009) and ATEC (Wong & Kit, 2010). This approach has been proven as an effective method to improve the performance of metrics, for those words in MT outputs having semantically equivalent counterparts in references can be appropriately rewarded. Nevertheless, such approach of identifying synonyms with WordNet may not be able to fully describe the similarity of words between MT outputs and references. WordNet has been argued for the granularity of sense distinctions which are too fine-grained (Navigli, 2006; Snow et al., 2007), that may cause the missing of some potential synonym pairs under a coarser standard from lay users. Furthermore, most metrics will consider an MT candidate word as unrelated when there is none of exact match or synonym found in references, this will lead to a reduction of the evaluation score. Indeed, we think that apart from the exact match and synonym match, word pairs in similar meaning should not be neglected in MT evaluation. What is needed is a measure of word similarity to find out these word pairs.

In this paper, we investigate the utilization of current word similarity measures in MT evaluation for finding out semantically similar word pairs to improve the performance of MT evaluation metrics. Those word similarity measures, both knowledge-based and corpus-based, have been widely applied in various NLP tasks in which their performance and reliability were proven. Their performance in MT evaluation, however, is still unknown, that will be our aim explored in the following experiments.

2. Semantic Similarity Measures

The formalization and quantification of lexical semantic similarity has been a problem in computational linguistics for many years. Different measures were proposed that rely on various kinds of resources and interpret the notion of semantic similarity in different manners. Previous researches (Budanitsky & Hirst, 2001,2006; Pucher, 2005; Liu et al., 2006) have attempted to compare these competing approaches to determine their validity, however, the results are rather inconsistent in terms of their correlation with human judgments. In general, it is suggested that the performance of these similarity measures is merely application-dependent, each of them may show different degree of merit depending on the context of use. In this study, eight different measures of semantic similarity are selected for the task of MT evaluation, including seven knowledge-based measures relying on WordNet as their knowledge source, plus one corpus-based measure trained with corpora.

The WordNet-based measures actually compute the similarity between two concepts (*synsets*) that the words in question belong to respectively. Some common notions shared by different measures include:

(i) the *length* which is the number of the fewest nodes between concepts c_1 and c_2;

(ii) the *depth* which is the *length* between concept c_1 and the global root node, i.e., *depth*(c_1) = *length*(*root*,c_1);

(iii) the *least common subsumer* (*lcs*) which is the most specific ancestor concept of both concepts c_1 and c_2;

(iv) the *information content* (*IC*) which is the specificity of a concept, measured by:

$$IC(c) = -\log p(c)$$

where $p(c)$ denotes the probability of the occurrence of concept c in a corpus.

The different similarity measures are then introduced as follows.

wup: Wu and Palmer (1994) measures similarity between concepts c1 and c2 in a hierarchy as:

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

lch: Leacock and Chodorow (1998) make use of the length between concepts to determine their similarity:

$$sim_{lch}(c_1, c_2) = -log\frac{length(c_1, c_2)}{2 \times \max depth(c)}$$

where $\max depth(c)$ refers to the maximum depth of a concept in the WordNet hierarchy.

res: Resnik's (1995) approach brings together a knowledge base and corpus statistics. The notion of similarity is defined as the extent to which two concepts share information in common, that is materialized as their least common subsumer. The measurement of similarity is then formulated as:

$$sim_{res}(c_1, c_2) = IC(lcs(c_1, c_2))$$

jcn: Jiang and Conrath's (1997) measure also utilizes the notion of information content. Their difference with Resnik's is the combination of both edge counts in WordNet and the information content of concepts:

$$sim_{jcn}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2))}$$

lin: Lin's (1998) similarity measure intends to be universally applicable to arbitrary objects, described by his theorem that "the similarity between *A* and *B* is measured by the ratio between the amount of information needed to state their commonality and the information needed to fully describe what they are". This is formulated into:

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

hso: Hirst and St-Onge (1998) conceives semantic similarity as the strength of semantic relationship between two concepts. This is represented by the *length* and the number of *direction changes* in the path connecting the concepts. Different relations between synsets in WordNet are classified into three directions including up, down and horizontal. The strength of semantic relationship is further categorized into extra-strong, strong, medium-strong and weak, where the first two categories will be given pre-defined similarity values. For medium-strong the value is calculated as follows:

$$sim_{hso}(c_1, c_2) = C - path\ length - k \times d$$

where d is the number of direction changes, and C and k are constants. The relationship is strong when a path is not too long and "does not change direction too often".

lesk: Banerjee and Pedersen's (2002) measure determines similarity according to the number of overlaps between the glosses of synsets that two concepts belong to. Formulated as follows:

$$sim_{lesk}(c_1, c_2) = \sum_{j \in S} \sum_{i \in O} overlap_{i,j}(g(c_1), g(c_2))^2$$

where

- $g(c)$ refers to the synset gloss of concept;
- $overlap(g_1, g_2)$ refers to the longest overlap between two glosses;
- O refers to all overlaps that can be matched;
- S refers to all related synsets of the concepts.

The length of the overlap contributes significantly to the score, a longer consecutive match is rewarded by the square of the number of its words in the match.

Apart from the above WordNet-based measures, a corpus-based measure, namely Latent Semantic Analysis (*LSA*) (Landauer et al., 1998) is also selected in our experiments. It is a kind of statistical computation to analyze the relationships between a set of documents and the words they contain. Its underlying assumption is that word meanings are mutually determined and constrained by their contextual information. The similarity between two words, therefore, can be accounted through analysis of their co-occurrence words in corpora.

The deployment of *LSA* involves the training of a semantic space that transforms text corpora into a mathematical representation. It is a matrix containing all unique word in corpora, word occurrence statistics, and weights of the word occurrence frequencies that represent the relative importance of a word in a particular text and the representativeness of this word in a domain of discourse. The matrix is then decomposed via singular value decomposition into three other matrices which are the product of the semantic space ready to be utilized. Every word in the semantic space can be represented by a multi-dimensional vector. The similarity of two words (w_1, w_2) is compared by the cosine of the angle between their vectors (v_1, v_2), where:

$$sim_{LSA}(w_1, w_2) = \frac{v_1 \cdot v_2}{\| v_1 \| * \| v_2 \|}$$

The application of *LSA* in measuring MT adequacy is explored in Reeder (2006). In that work it is used as a primary approach to evaluate MT outputs in the granularity of system, document and paragraph levels. The results are positive in terms of correlation with human judgments, but not as good as *LSA* is used in grading

human essays. In our experiments, LSA is treated as an assistance of other evaluation metrics for measurement of semantic similarity of words only.

3. Experiments

The experiments focus on two main questions. First, for each semantic measure described in the previous section, we want to know the degree of similarity that a word pair from an MT output and a reference translation should have in order to contribute to the quality of an MT output. Second, how much performance gain an MT evaluation metric can benefit from these semantic similarity measures.

3.1 Setting

The MetricsMATR08 development data (Przybocki et al., 2009) is adopted in our experiments. It consists of 1992 outputs from eight different MT systems with human assessments and four versions of reference translation.

WordNet 2.1 is used for those knowledge-based measures. A pre-compiled *LSA* semantic space[1] trained with texts in general domain at college level is selected.

The semantic similarity measures are integrated with a fundamental MT evaluation metric based on unigram matches between an MT output and its reference translation. A unigram match can be an exact word, a synonym or a semantically similar word, all kinds of match carry the same weight. This ensures that the metric is sensitive to word choice only, and disregards all other features such as word order or syntax. All the word pairs retrieved for similarity measurement are verified for their existence in both WordNet and the *LSA* semantic space, as well as the same part-of-speech, to ensure that the numbers of word pairs for every similarity measure are equal. In practice, the evaluation metric is divided into the precision (*p*) and recall (*r*) between the number of unigram matches and the length of the MT output (*c*) and reference translation (*t*) respectively, and their harmonic F-measure (*f*), formulated as follows.

$$p(c,t) = \frac{match(c,t)}{length(c)} \quad r(c,t) = \frac{match(c,t)}{length(r)} \quad f(c,t) = \frac{2pr}{p+r}$$

This unigram-based metric is taken as the basis of the design of many more advanced MT evaluation metrics, such as the precision oriented metric like BLEU (1-gram), recall oriented like METEOR, and F-measure oriented like ATEC. The experiment results in this setting are therefore representable for different kinds of evaluation metrics in use.

3.2 Results

A fundamental question to identify semantically similar word pairs is the definition of the degree of similarity. This is evaluated by testing each similarity measure via a hill climbing method to seek its optimal similarity threshold, such that the similarity value of a word pair has to be above the threshold in order to be considered as semantically close enough. Table 1a shows the optimal

Metric	Reference	*jcn*	*lin*	*lesk*	*res*	*hso*	*lch*	*wup*	*LSA*
precision	multiple	0.46	0.89	4573	10.91	16	2.94	0.96	0.69
	single	0.46	0.89	11611	10.91	13.75	2.94	0.96	0.70
recall	multiple	0.46	0.89	5112	10.91	16	2.94	0.96	0.68
	single	0.46	0.89	11084	10.91	13.75	2.94	0.96	0.71
F-measure	multiple	0.46	0.89	5112	10.91	16	2.94	0.96	0.68
	single	0.46	0.89	7513	10.91	13.75	2.94	0.96	0.68

Table 1a. Optimal thresholds of each similarity measure

Metric	Reference	*jcn*	*lin*	*lesk*	*res*	*hso*	*lch*	*wup*	*LSA*	exact
precision	multiple	.5639	.5668	.5665	.5658	.5679	.5703	.5706	.5720	.5666
	single	.4524	.4538	.4558	.4570	.4583	.4583	.4595	.4617	.4564
recall	multiple	.6049	.6067	.6099	.6100	.6081	.6120	.6094	.6120	.6095
	Single	.5278	.5290	.5303	.5320	.5325	.5336	.5337	.5362	.5308
F-measure	multiple	.6236	.6260	.6263	.6267	.6272	.6305	.6295	.6325	.6261
	single	.5202	.5216	.5223	.5242	.5260	.5259	.5271	.5307	.5228

Table 1b. Correlations of each similarity measure under optimal thresholds

Metric	Reference	*jcn*	*lin*	*lesk*	*res*	*hso*	*lch*	*wup*	*LSA*
precision	multiple	-0.48%	0.04%	-0.02%	-0.13%	0.22%	0.66%	0.71%	0.94%
	single	-0.87%	-0.56%	-0.13%	0.13%	0.42%	0.42%	0.68%	1.16%
recall	multiple	-0.76%	-0.46%	0.07%	0.08%	-0.23%	0.41%	-0.02%	0.41%
	Single	-0.57%	-0.34%	-0.09%	0.22%	0.32%	0.52%	0.54%	1.01%
F-measure	multiple	-0.40%	-0.02%	0.03%	0.09%	0.17%	0.70%	0.54%	1.02%
	single	-0.50%	-0.24%	-0.09%	0.25%	0.61%	0.59%	0.81%	1.50%

Table 1c. Percentage changes of correlation of each similarity measure compared with exact match

[1] http://lsa.colorado.edu/

	precision				recall				F-measure			
	single		multiple		single		multiple		single		multiple	
exact	.6646		.7971		.6529		.7790		.6543		.7840	
synonyms	.6836	2.86%	.8116	1.82%	.6715	2.85%	.7923	1.71%	.6730	2.86%	.7978	1.76%
wup	.6766	1.81%	.8083	1.41%	.6612	1.27%	.7893	1.32%	.6662	1.82%	.7947	1.36%
LSA	.6775	1.94%	.8076	1.32%	.6656	1.95%	.7889	1.27%	.6670	1.94%	.7941	1.29%
wup & LSA	.6853	3.11%	.8142	2.15%	.6732	3.11%	.7950	2.05%	.6747	3.12%	.8005	2.10%

Table 2. Average evaluation scores of different MT evaluation measures

	precision				recall				F-measure			
	single		multiple		single		multiple		single		multiple	
exact	.4564		.5666		.5308		.6095		.5228		.6261	
synonyms	.4597	0.74%	.5651	-0.26%	.5352	0.82%	.6068	-0.45%	.5286	1.10%	.6265	0.06%
wup	.4594	0.65%	.5705	0.70%	.5335	0.52%	.6094	-0.02%	.5270	0.80%	.6295	0.54%
LSA	.4596	0.71%	.5715	0.86%	.5349	0.77%	.6118	0.37%	.5291	1.19%	.6321	0.96%
wup & LSA	.4612	1.05%	.5725	1.05%	.5365	1.09%	.6103	0.12%	.5319	1.73%	.6332	1.13%

Table 3. Correlations of different MT evaluation measures

similarity thresholds of each similarity measure applied in the three MT evaluation metrics using multiple or single reference translation, that result in the highest correlation with human assessments. For most similarity measures, their optimal thresholds are rather consistent under different settings, except *lesk* because it is largely determined by the number of words in synset glosses which varies for different words. Their corresponding correlation values, measured by Pearson correlation coefficient at segment level, are shown in Table 1b, the correlations of the metrics using exact match only are listed for reference as well. Table 1c shows the percentage changes of correlation of each similarity measure compared with exact match. It is shown that, unexpectedly, not all similarity measures contribute positively to the evaluation metrics. Measures like *jcn*, *lin* and *lesk* even lead to degradation of metric performance. On the other hand, *lch*, *wup* and *LSA* are better measures in this experiment, where *LSA* gives the best performance in all different settings.

Instead of solely utilizing *LSA* as the only similarity measure to supplement an evaluation metric, however, we think that the hybrid use of both WordNet-based similarity and *LSA* is a better alternative. As they rely on different resources, their similar word sets may be able to complement each other. We select *wup* to further evaluate this idea, for the noticeable correlation gain it brings to the metric among all similarity measures, and also for its value interval which is between 0 and 1, and therefore more accountable.

Table 2 and 3 show the average scores and correlations of the evaluation metrics in various settings. The exact match serves as a baseline and the WordNet synonym match is provided here for comparison. The similarity measures *wup* and *LSA* are tested alone as well as together. The percentages refer to the changes of evaluation scores and correlations of the evaluation metrics with the aid of synonym match or word similarity measures, compared with exact match. It shows that the use of *wup* or *LSA* both allows more matches than exact match only, as reflected in the raises of precision, recall and F-measure in both single and multiple reference settings. Such increases of evaluation scores come together with an observable improvement in correlations. Furthermore, the combination of the two similarity measures results in the highest evaluation scores in all settings. This verifies our preceding notion that the semantically similar words retrieved by *wup* and *LSA* are complementary. From another point of view, this also reveals how many words that should be considered in MT evaluation have been neglected by current evaluation metrics. As shown in the correlations, the contribution of similarity measures outperforms synonym match, in most settings the correlation gains are higher than 1%.

4. Conclusion

We have focused on the problem of current MT evaluation metrics that semantically similar word pairs are disregarded in the comparison of MT outputs and reference translations, such problem would lead to an underestimation of the quality of certain MT outputs. Our experiments of word similarity measures have shown that two of them, i.e., *wup* and *LSA*, are better in identifying word pairs in close meaning for MT evaluation.

Following this line of research, our current work continues to explore the possibilities and weaknesses of word similarity measures. In particular, some of them, in principle, assess the semantic relatedness of words rather than their similarity. For example, a word pair 'committee' and 'chairman' gets a high value in *LSA* but they are indeed not very close in meaning. Besides, most WordNet similarity measures only work on nouns and verbs as restricted by the structure of WordNet. The effect of these inadequacies on MT evaluation has to be investigated. On the other hand, we have showed that the combination of multiple similarity measures generates a better performance. As each similarity measure may have its own strength on particular word types, their subsequence exploration may reveal a new way to dynamically opt for a suitable one for a specific group of words.

5. Acknowledgements

The work described in this paper is supported by City University of Hong Kong through the Strategic Research Grant (SRG) 7002267.

6. References

Babych, B. & Hartley, A. (2008). Sensitivity of Automated MT Evaluation Metrics on Higher Quality MT Output: BLEU vs Task-Based Evaluation Methods. *The Sixth International Language Resources and Evaluation (LREC'08)*.

Banerjee, S. & Lavie, A. (2005). METEOR: an Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *ACL-2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, University of Michigan, Ann Arbor, pages 65-72.

Banerjee, S. & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-02)*. Mexico City.

Budanitsky, A. & Hirst, G. (2001). Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*. Pittsburgh.

Budanitsky, A. & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. 32(1): 13-47.

Callison-Burch, C., Osborne, M. & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *11th Conference of the European Chapter of the Association for Computational Linguistics*. pages 249–256.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136-158.

Chan, Y.S. & Ng, H.T. (2008). MAXSIM: a Maximum Similarity Metric for Machine Translation Evaluation. In *Proceedings of ACL-08:HLT*, pages 55-62.

Hirst, G. & St-Onge, D. (1998). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Christiane Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. MIT Press, pages 305-332.

Jiang, J. J. & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*. Taiwan.

Landauer, T. K., Foltz, P. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* 25.

Leacock, C. & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Christiane Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. MIT Press, pages 265-283.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Madison, WI.

Liu, P-Y., Zhao, T-J. & Yu, X-F. (2006). Application-Oriented Comparison and Evaluation of Six Semantic Similarity Measures Based on WordNet. In *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*. Dalian, pages 2605-2610.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, pages 105-112.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report*.

Przybocki, M., Peterson, K. & Bronsart, S. (2009). 2008 NIST Metrics for Machine Translation (MetricsMATR08) Development Data. Linguistic Data Consortium, Philadelphia.

Pucher, M. (2005). Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech. In *Proceedings of the International Workshop on Computational Semantics*. Tilburg, Netherlands.

Reeder, F. (2006). Measuring MT Adequacy Using Latent Semantic Analysis. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*. Cambridge, Massachusetts, pages 176-184.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, pages 448-453.

Snover, M., Madnani, N., Dorr, B. & Schwartz, R. (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Snow, R., Prakash, S., Jurafsky, D. & Ng, A.Y. (2007). Learning to Merge Word Senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, pages 1005-1014.

Wong, B. & Kit, C. (2010). ATEC: Automatic Evaluation of Machine Translation via Word Choice and Word Order. *Machine Translation*, 23(2):141-155.

Wu, Z. & Palmer, M. (1994). Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, New Mexico.