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 instance, 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 c₁ and c₂;
(ii) the depth which is the length between concept c₁ and the global root node, i.e.,\( depth(c_1) = \text{length(root,}c_1) \);
(iii) the least common subsumer (lcs) which is the most specific ancestor concept of both concepts c₁ and c₂;
(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.

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

\[
\text{sim}_{\text{wup}}(c_1, c_2) = \frac{2 \times \text{depth(lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}
\]

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

\[
\text{sim}_{\text{lch}}(c_1, c_2) = -\log\frac{\text{length}(c_1, c_2)}{2 \times \text{max depth}(c)}
\]

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

\textit{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:

\[
\text{sim}_{\text{res}}(c_1, c_2) = IC(lcs(c_1, c_2))
\]

\textit{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:

\[
\text{sim}_{\text{jcn}}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2))}
\]

\textit{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:

\[
\text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}
\]

\textit{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:

\[
\text{sim}_{\text{hso}}(c_1, c_2) = C - \text{path length} - k \times d
\]

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”.

\textit{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:

\[
\text{sim}_{\text{lesk}}(c_1, c_2) = \sum_{j \in S} \sum_{i \in O} \text{overlap}_{ij}(g(c_1), g(c_2))^2
\]

where

- \( g(c) \) refers to the synset gloss of concept;
- \( \text{overlap}_{ij}(g_1, g_2) \) 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:

\[
\text{sim}_{\text{LSA}}(w_1, w_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\|\overrightarrow{v_1}\| \|\overrightarrow{v_2}\|}
\]

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 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Reference</th>
<th>jcn</th>
<th>lin</th>
<th>lesk</th>
<th>res</th>
<th>hso</th>
<th>lch</th>
<th>wup</th>
<th>LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>multiple</td>
<td>0.46</td>
<td>0.89</td>
<td>4573</td>
<td>10.91</td>
<td>16</td>
<td>2.94</td>
<td>0.96</td>
<td>0.69</td>
</tr>
<tr>
<td>recall</td>
<td>single</td>
<td>0.46</td>
<td>0.89</td>
<td>11611</td>
<td>10.91</td>
<td>13.75</td>
<td>2.94</td>
<td>0.96</td>
<td>0.70</td>
</tr>
<tr>
<td>F-measure</td>
<td>multiple</td>
<td>0.46</td>
<td>0.89</td>
<td>5112</td>
<td>10.91</td>
<td>16</td>
<td>2.94</td>
<td>0.96</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>0.46</td>
<td>0.89</td>
<td>11084</td>
<td>10.91</td>
<td>13.75</td>
<td>2.94</td>
<td>0.96</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 1a. Optimal thresholds of each similarity measure

<table>
<thead>
<tr>
<th>Metric</th>
<th>Reference</th>
<th>jcn</th>
<th>lin</th>
<th>lesk</th>
<th>res</th>
<th>hso</th>
<th>lch</th>
<th>wup</th>
<th>LSA</th>
<th>exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>multiple</td>
<td>0.5639</td>
<td>0.5668</td>
<td>0.5665</td>
<td>0.5658</td>
<td>0.5679</td>
<td>0.5703</td>
<td>0.5706</td>
<td>0.5720</td>
<td>0.5666</td>
</tr>
<tr>
<td>recall</td>
<td>single</td>
<td>0.4524</td>
<td>0.4538</td>
<td>0.4558</td>
<td>0.4570</td>
<td>0.4583</td>
<td>0.4583</td>
<td>0.4595</td>
<td>0.4617</td>
<td>0.4654</td>
</tr>
<tr>
<td>F-measure</td>
<td>multiple</td>
<td>0.6049</td>
<td>0.6067</td>
<td>0.6099</td>
<td>0.6100</td>
<td>0.6081</td>
<td>0.6120</td>
<td>0.6094</td>
<td>0.6120</td>
<td>0.6095</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>0.5278</td>
<td>0.5290</td>
<td>0.5303</td>
<td>0.5320</td>
<td>0.5325</td>
<td>0.5335</td>
<td>0.5336</td>
<td>0.5337</td>
<td>0.5362</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Reference</th>
<th>jcn</th>
<th>lin</th>
<th>lesk</th>
<th>res</th>
<th>hso</th>
<th>lch</th>
<th>wup</th>
<th>LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>multiple</td>
<td>-0.48%</td>
<td>0.04%</td>
<td>-0.02%</td>
<td>-0.13%</td>
<td>0.22%</td>
<td>0.66%</td>
<td>0.71%</td>
<td>0.94%</td>
</tr>
<tr>
<td>recall</td>
<td>single</td>
<td>-0.87%</td>
<td>-0.56%</td>
<td>-0.13%</td>
<td>0.13%</td>
<td>0.42%</td>
<td>0.42%</td>
<td>0.68%</td>
<td>1.16%</td>
</tr>
<tr>
<td>F-measure</td>
<td>multiple</td>
<td>-0.76%</td>
<td>-0.46%</td>
<td>0.07%</td>
<td>0.08%</td>
<td>-0.23%</td>
<td>-0.41%</td>
<td>-0.02%</td>
<td>0.41%</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>-0.57%</td>
<td>-0.34%</td>
<td>-0.09%</td>
<td>0.22%</td>
<td>0.32%</td>
<td>0.52%</td>
<td>0.54%</td>
<td>1.01%</td>
</tr>
</tbody>
</table>

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

1. http://lsa.colorado.edu/
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 synom 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 selection 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 selection...
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