LAYERED: Metric for Machine Translation Evaluation

Shubham Gautam
Computer Science & Engineering,
IIT Bombay
shubhamg@cse.iitb.ac.in

Pushpak Bhattacharyya
Computer Science & Engineering,
IIT Bombay
pb@cse.iitb.ac.in

Abstract

This paper describes the LAYERED metric which is used for the shared WMT'14 metrics task. Various metrics exist for MT evaluation: BLEU (Papineni, 2002), METEOR (Alon Lavie, 2007), TER (Snover, 2006) etc., but are found inadequate in quite a few language settings like, for example, in case of free word order languages. In this paper, we propose an MT evaluation scheme that is based on the NLP layers: lexical, syntactic and semantic. We contend that higher layer metrics are after all needed. Results are presented on the corpora of ACL-WMT, 2013 and 2014. We end with a metric which is composed of weighted metrics at individual layers, which correlates very well with human judgment.

1 Introduction

Evaluation is an integral component of machine translation (MT). Human evaluation is difficult and time consuming so there is a need for a metric which can give the better evaluation in correlation to human judgement. There are several existing metrics such as: BLEU, METEOR *etc.* but these only deal with the lexical layer combining *bag of words* and *n-gram* based approach.

We present an analysis of BLEU and the higher layer metrics on the ACL WMT 2013 corpora with 3 language pairs: French-English, Spanish-English and German-English. For syntactic layer, we considered three metrics: Hamming score, Kendall's Tau distance score and the spearman rank score. Syntactic layer metrics take care of reordering within the words of the sentences so these may play an important role when there is a decision to be made between two MT output sentences of two different systems when both the sentences have same number of n-gram matches *wrt* the reference sentence but there is a difference in the ordering of the sentence. We will discuss these metrics in detail in the following sections. The next NLP layer in consideration is the semantic layer which deals with the meaning of the sentences. For semantic layer, we considered two metrics: Shallow semantic score and Deep semantic score. On semantic layer, we considered entailment based measures to get the score.

Ananthkrishnan et al. (2007) mentioned some issues in automatic evaluation using BLEU. There are some disadvantages of the existing metrics also such as: *BLEU* does not take care of reordering of the words in the sentence. *BLEU*-like metrics can give same score by permuting word order. These metrics can be unreliable at the level of individual sentences because there can be small number of n-grams involved. We would see in this paper that the correlation of BLEU is lower compared to the semantic layer metrics.

Section 2 presents the importance of each NLP layer in evaluation of MT output. It discusses the metrics that each layer contributes to the achievement of the final result. In section 3, various experiments are presented with each metric on the top 10 ranking systems of WMT 13 corpora which are ranked on the basis of the human ranking. Each metric is discussed with the graphical representation so that it would become clear to analyze the effect of each metric. In section 4, spearman correlation of the metrics is calculated with human judgement and comparisons are shown. In section 5, we discuss the need of a metric which should be a combination of the metrics presented in the above sections and present a weighted metric which is the amalgamation of the metrics at individual layers.

2 Related Work

Machine translation evaluation has always remained as the most popular measure to judge the quality of a system output compared to the reference translation. Papineni (2002) proposed BLEU as an automatic MT evaluation metric which is based on the n-gram matching of the reference and candidate sentences. This is still considered as the most reliable metric and used widely in the MT community for the determination of the translation quality. BLEU averages the precision for unigram, bigram and up to 4-gram and applies a length penalty if the generated sentence is shorter than the best matching (in length) reference translation. Alternative approaches have been designed to address problems with BLEU. Doddington and George (2003) proposed NIST metric which is derived from the BLEU evaluation criterion but differs in one fundamental aspect: instead of n-gram precision, the information gain from each n-gram is taken into account. TER (Snover, 2006) tries to improve the hypothesis/reference matching process based on the edit-distance and METEOR (Alon Lavie, 2007) considered linguistic evidence, mostly lexical similarity, for more intelligent matching. Liu and Gildea (2005), Owczarzak et al. (2007), and Zhang et al. (2004) use syntactic overlap to calculate the similarity between the hypothesis and the reference. Padó and Galley (2009) proposed a metric that evaluates MT output based on a rich set of textual entailment features. There are different works that have been done at various NLP layers. Giménez tl al. (2010) provided various linguistic measures for MT evaluation at different NLP layers. Ding Liu and Daniel Gildea (2005) focussed the study on the syntactic features that can be helpful while evaluation.

3 Significance of NLP Layers in MT Evaluation

In this section, we discuss the different NLP layers and how these are important for evalution of MT output. We discuss here the significance of three NLP layers: Lexical, Syntactic and Semantic layers.

3.1 Lexical Layer

Lexical layer emphasizes on the comparison of the words in its original form irrespective of any lexical corpora or any other resource. There are some metrics in MT evaluation which considers only these features. Most popular of them is *BLEU*, this is based on the n-gram approach and considers the matching upto 4-grams in the reference and the candidate translation. BLEU is designed to approximate human judgement at a corpus level, and performs badly if used to evaluate the quality of individual sentences. Another important metric at this layer is TER (Translation Edit Rate) which measures the number of edits required to change a system output into one of the references. For our experiments, we would consider BLEU as the baseline metric on lexical layer.

3.2 Syntactic Layer

Syntactic layer takes care of the syntax of the sentence. It mainly focusses on the reordering of the words within a sentence. Birch and Osborne (2011) has mentioned some metrics on this layer: Hamming score and Kendall's Tau Distance (KTD) score. We additionally calculated the spearman rank score on this layer. Scores are calculated first by giving ranking of words in the reference sentence and then putting the rank number of the word in the candidate sentence. Now, we have the relative ranking of the words of both the sentences, so final score is calculated.

3.3 Semantic Layer

Semantic layer goes into the meaning of the sentence, so we need to compare the dependency tree of the sentences. At this layer, we used *entailment* based metrics for the comparison of dependencies. Padó and Galley (2009) illustrated the use of text entailment based features for MT evaluation. We introduced two metrics at this layer: *first* is *Shallow semantic score*, which is based on the dependencies generated by a shallow parser and then the dependency comparison is carried out. *Second* is *Deep semantic score*, which goes more deep into the semantic of the sentence. For *shallow semantic score*, we used stanford dependency parser (Marie-Catherine et al., 2006) while for *deep semantic score*, we used UNL (Universal Networking Language)[1] dependency generator.

Semantic layer may play an important role when there are different words in two sentences but they are synonym of each other or are related to each other in some manner. In this case, lexical and syntactic layers can't identify about the simi-

[1] http://www.undl.org/unlsys/unl/unl2005/UW.htm

larity of the sentences because there exist a need of some semantic background knowledge which occurs at the semantic layer. Another important role of semantic layer is that there can be cases when there is reordering of the phrases in the sentences, *e.g.,* active-passive voice sentences. In these cases, dependencies between the words remain intact and this can be captured through dependency tree generated by the parser.

4 Experiments

We conducted the experiments on WMT 13 corpora for French-English, Spanish-English and German-English language pairs. We calculated the score of each metric for the top 10 ranking system (wmt, 2013) (as per human judgement) for each language pair.

Note:

1. In the graphs, metric score is multiplied by 100 so that a better view can be captured.

2. In each graph, the scores of French-English (fr-en), Spanish-English (es-en) and German-English (de-en) language pairs are represented by red, black and blue lines respectively.

4.1 BLEU Score

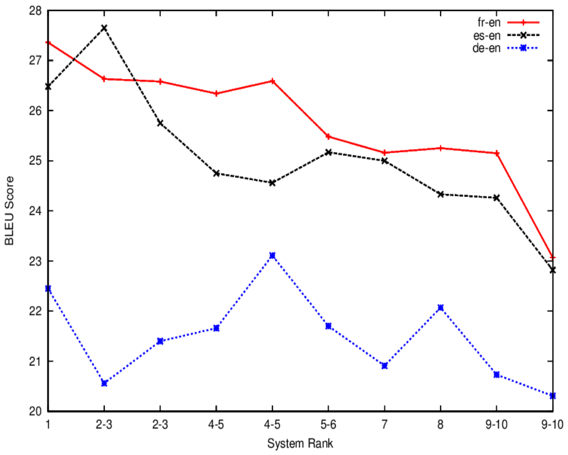

Figure 1: BLEU Score

We can see from the graph of fig. 1 that for de-en and es-en language pair, BLEU is not able to capture the phenomenon appropriately. In fact, it is worse in de-en pair. Because the graph should be of decreasing manner *i.e.,* as the rank of the system increases (system gets lower rank compared to the previous one), the score should also decrease.

4.2 Syntactic Layer

Because the BLEU score was not able to capture the idealistic curve in the last section so we considered the syntactic layer metrics. This layer is considered because it takes care of the reordering of the words within the sentence pair. The idea here is that if one candidate translation has lower reordering of words *w.r.t.* reference translation then it has higher chances of matching to the reference sentence.

4.2.1 Hamming Score

The hamming distance measures the number of disagreements between two permutations. It is formulated as follows:

$$d_h(\pi,\sigma) = 1 - \frac{\sum_{i=1}^{n} x_i}{n}, x_i = \begin{cases} 0; \; if \; \pi(i) = \sigma(i) \\ 1; \; otherwise \end{cases}$$

where, n is the length of the permutation.

Hamming scores for the all three language pairs mentioned above are shown in fig. 2. As we can see from the graph that initially its not good for the top ranking systems but it follows the ideal curve for the discrimination of lower ranking systems for the language pairs.

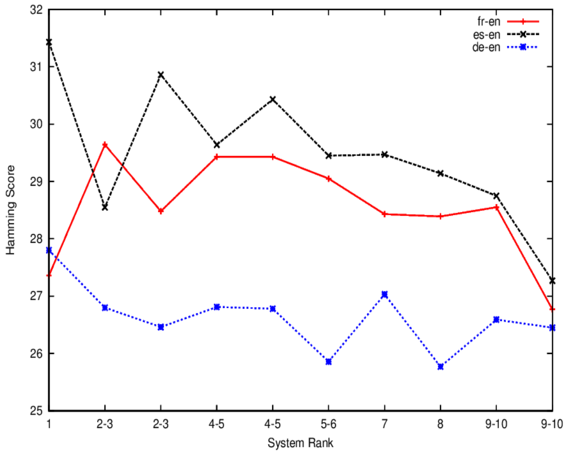

Figure 2: Hamming Score

4.2.2 Kendall's Tau Distance (KTD)

Kendall's tau distance is the minimum number of transpositions of two adjacent symbols necessary to transform one permutation into another. It represents the percentage of pairs of elements which share the same order between two permutations. It is defined as follows:

$$d_k(\pi, \sigma) = 1 - \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}}{Z}}$$

$$where, z_{ij} = \begin{cases} 0; \; if\ \pi(i) < \pi(j)\ and\ \sigma(i) < \sigma(j) \\ 1; \; otherwise \end{cases}$$

This can be used for measuring word order differences as the relative ordering words is taken into account. KTD scores are shown in fig. 3. It also follows the same phenomenon as the hamming score for fr-en and es-en pair but for de-en pair, it gives the worst results.

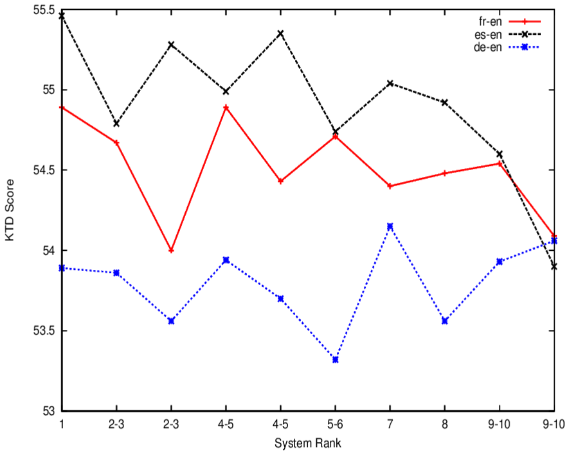

Figure 3: KTD Score

4.2.3 Spearman Score

Spearman rank correlation coefficient is basically used for assessing how well the relationship between two variables can be described using a monotonic function. Because we are using syntactic layer metrics to keep track of the reordering between two sentences, so this can be used by ranking the words of the first sentence (ranging from 1 to n, where n is the length of the sentence) and then checking where the particular word (with index i) is present in the second sentence in terms of ranking. Finally, we calculated the spearman score as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where, $d_i = x_i - y_i$ is the difference between the ranks of words of two sentences.

Spearman score lies between -1 to +1 so we convert it to the range of 0 to +1 so that all the metrics would lie in the same range.

4.3 Semantic Layer

We can see from the last two sections that there were some loopholes on the metrics of both the layers as can be seen in the graphical representations. So, there arises a need to go higher in the hierarchy. The next one in the queue is semantic layer which takes care of the meaning of the sentences. At this layer, we considered two metrics. Both metrics are based on the concept of *text entailment*. First we should understand, what is it?

Text Entailment

According to wikipedia[2], "Textual entailment (TE) in natural language processing is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. In the TE framework, the entailing and entailed texts are termed *text* (t) and *hypothesis* (h), respectively."

First, the dependencies for both reference (R) as well as candidate (C) translation are generated using the parser that is used (will vary in both the following metrics). Then, the entailment phenomenon is applied from R to C *i.e.,* dependencies of C are searched in the dependency graph of R. Matching number of dependencies are calculated, then a score is obtained as follows:

$$Score_{R-C} = \frac{No.\ of\ matched\ dependencies\ of\ C\ in\ R}{Total\ no.\ of\ dependencies\ of\ C} \tag{1}$$

Similarly, another score is also obtained by applying the entailment phenomenon in the reversed direction *i.e. from C to R* as follows:

$$Score_{C-R} = \frac{No.\ of\ matched\ dependencies\ of\ R\ in\ C}{Total\ no.\ of\ dependencies\ of\ R} \tag{2}$$

Final score is obtained by taking the average of the above two scores as follows:

$$Score_{final} = \frac{Score_{R-C} + Score_{C-R}}{2} \tag{3}$$

Now, we discuss how can we use this concept in the metrics at semantic layer:

4.3.1 Shallow Semantic Score

This metric uses the stanford dependency parser[3] to generate the dependencies. After getting the dependencies for both reference (R) as well as candidate (C) translation, entailment phenomenon is applied and the final score is obtained using eq. (3).

[2] http://wikipedia.org/

[3] http://nlp.stanford.edu/software/lex-parser.shtml

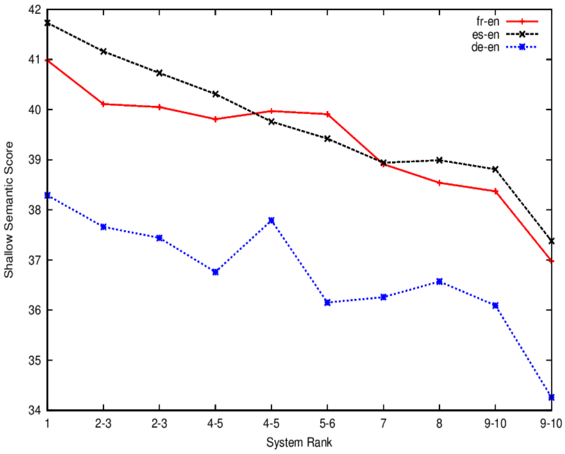

Figure 4: Shallow Semantic Score

We can see from fig. 4 that for French-English and Spanish-English pairs, the graph is very good compared to the other metrics at the lower layers. In fact, there is only one score in es-en pair that a lower ranking system gets better score than the higher ranking system.

4.3.2 Deep Semantic Score

This metric uses the UNL dependency graph generator for taking care of the semantic of the sentence that shallow dependency generator is not able to capture. Similar to the shallow semantic score, after getting the dependencies from the UNL, entailment score is calculated in both directions *i.e.* $R \rightarrow C$ *and* $C \rightarrow R$.

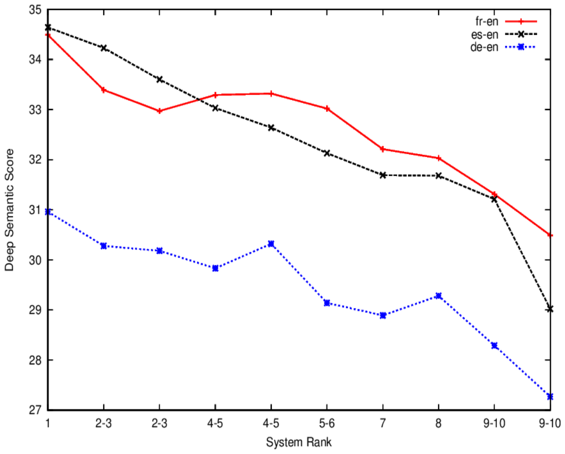

Figure 5: Deep Semantic Score

Fig. 5 shows that deep semantic score curve also follows the same path as shallow semantic score. In fact, for Spanish-English pair, the path is ideal *i.e., the score is decreasing as the system rank is increasing*.

5 Correlation with Human Judgement

We calculated spearman rank correlation coefficient for the different scores calculated in the last section. This score ranges from -1 to +1. Form table 1, we can see that the correlation score is better with semantic layer metrics compared to the BLEU score (lower layer metrics). In comparison to the WMT 13 results (wmt-result, 2013), $\rho_{Shallow}$ score for French-English pair is intermediate between the highest and lowest correlation system. ρ_{Deep} score for Spanish-English is highest among all the systems presented at WMT 13. So, it arises a need to take into account the semantic of the sentence while evaluating the MT output.

Language Pair	ρ**BLEU**	ρ**Shallow**	ρ**Deep**
French-English	0.95	**0.96**	0.92
Spanish-English	0.89	0.98	**1.00**
German-English	0.36	0.88	0.89

Table 1: Correlation with BLEU Score, Shallow Semantic Score and Deep Semantic Score

6 Hybrid Approach

We reached to a situation where we can't ignore the score of any layer's metric because each metric helps to capture some of the phenomenon which other may not capture. So, we used a hybrid approach where the final score of our proposed metric depends on the layered metrics. As already said, we performed our experiments on ACL-WMT 2013 corpora, but it provided only the rank of the systems. Due to availability of ranking of the systems, we used SVM-rank to learn the parameters.

Our final metric looks as follows:
Final-Score = a*BLEU + b*Hamming + c*KTD + d*Spearman + e*Shallow-Semantic-Score + f*Deep-Semantic-Score

where, a,b,c,d,e,f are parameters

6.1 SVM-rank

SVM-rank learns the parameters from the training data and builds a model which contains the learned parameters. These parameters (model) can be used for ranking of a new set of data.

Parameters

We made the training data of the French-English, Spanish-English and German-English language

Metric	Pearson Correlation					
	fr-en	**de-en**	**hi-en**	**cs-en**	**ru-en**	**Average**
LAYERED	.973	.893	.976	.940	.843	.925
BLEU	.952	.831	.956	.908	.774	.884
METEOR	.975	.926	.457	.980	.792	.826
NIST	.955	.810	.783	.983	.785	.863
TER	.952	.774	.618	.977	.796	.823

Table 2: Correlation with different metrics in WMT 14 Results

pairs. Then we ran SVM-rank and obtained the scores for the parameters.

So, our final proposed metric looks like:
Final-Score = 0.26*BLEU + 0.13*Hamming + 0.03*KTD + 0.04*Spearman + 0.28*Shallow-Semantic-Score + 0.26* Deep-Semantic-Score

7 Performance in WMT 2014

Table 2 shows the performance of our metric on WMT 2014 data (wmt-result, 2014). It performed very well in almost all language pairs and it gave the highest correlation with human in Hindi-English language pair. On an average, our correlation was 0.925 with human considering all the language pairs. This way, we stood out on second position considering the average score while the first ranking system obtained the correlation of 0.942. Its clear from table 2 that the proposed metric gives the correlation better than the standard metrics in most of the cases. If we look at the average score of the metrics in table 2 then we can see that LAYERED obtains much higher score than the other metrics.

8 Conclusion

Machine Translation Evaluation is an exciting field that is attracting the researchers from the past few years and the work in this field is enormous. We started with the need of using higher layer metrics while evaluating the MT output. We understand that it might be a little time consuming but its efficient and correlation with human judgement is better with semantic layer metric compared to the lexical layer metric. Because, each layer captures some linguistic phenomenon so we can't completely ignore the metrics at individual layers. It gives rise to a hybrid approach which gives the weightage for each metric for the calculation of final score. We can see from the results of WMT 2014 that the correlation with LAYERED metric is better than the standard existing metrics in most of the language pairs.

References

Alexandra Birch, School of Informatics, University of Edinburgh *Reordering Metrics for Statistical Machine Translation.* Phd Thesis, 2011.

Alexandra Birch and Miles Osborne *Reordering Metrics for MT.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, series = HLT 2011.

Alon Lavie and Abhaya Agarwal. *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*, Proceedings of the Second Workshop on Statistical Machine Translation, StatMT 2007.

Ananthakrishnan R and Pushpak Bhattacharyya and M Sasikumar and Ritesh M Shah *Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU.* ICON, 2007.

Doddington and George *Automatic evaluation of machine translation quality using N-gram co-occurrence statistics, NIST.* Proceedings of the 2nd International Conference on Human Language Technology Research HLT 2002.

Ding Liu and Daniel Gildea *Syntactic Features for Evaluation of Machine Translation.* Workshop On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/or Summarization, 2005.

Findings of the 2013 Workshop on Statistical Machine Translation. ACL-WMT 2013.

Giménez, Jesús and Màrquez, Lluís *Linguistic Measures for Automatic Machine Translation Evaluation.* Machine Translation, December, 2010.

Liu D, Gildea D *Syntactic features for evaluation of machine translation.* ACL 2005 workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.

Owczarzak K, Genabith J, Way A *Evaluating machine translation with LFG dependencies.* Machine Translation 21(2):95119.

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. *Generating Typed Dependency Parses from Phrase Structure Parses*. LREC 2006.

Matthew Snover and Bonnie Dorr and Richard Schwartz and Linnea Micciulla and John Makhoul. *A Study of Translation Edit Rate with Targeted Human Annotation*, In Proceedings of Association for Machine Translation in the Americas, 2006.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. *BLEU: A Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002.

Results of the WMT13 Metrics Shared Task. ACL-WMT 2013.

Results of the WMT14 Metrics Shared Task. ACL-WMT 2014.

Sebastian Padó and Michel Galley and Dan Jurafsky and Chris Manning *Robust Machine Translation Evaluation with Entailment Features*. Proceedings of ACL-IJCNLP 2009, ACL 2009.

Zhang Y, Vogel S, Waibel A *Interpreting Bleu/NIST scores: how much improvement do we need to have a better system?*. In: Proceedings of the 4th international conference on language resources and evaluation. Lisbon, Portugal.