Sentence Level Machine Translation Evaluation as a Ranking Problem: one step aside from BLEU

Yang Ye
University of Michigan
yye@umich.edu

Ming Zhou
Microsoft Research Asia
mingzhou@microsoft.com

Chin-Yew Lin
Microsoft Research Asia
cyl@microsoft.com

Abstract

The paper proposes formulating MT evaluation as a ranking problem, as is often done in the practice of assessment by human. Under the ranking scenario, the study also investigates the relative utility of several features. The results show greater correlation with human assessment at the sentence level, even when using an n-gram match score as a baseline feature. The feature contributing the most to the rank order correlation between automatic ranking and human assessment was the dependency structure relation rather than BLEU score and reference language model feature.

1 Introduction

In recent decades, alongside the growing research on Machine Translation (MT), automatic MT evaluation has become a critical problem for MT system developers, who are interested in quick turnaround development cycles. The state-of-the-art automatic MT evaluation is an n-gram based metric represented by BLEU (Papineni et al., 2001) and its variants. Ever since its creation, the BLEU score has been the gauge of Machine Translation system evaluation. Nevertheless, the research community has been largely aware of the deficiency of the BLEU metric. BLEU captures only a single dimension of the vitality of natural languages: a candidate translation gets acknowledged only if it uses exactly the same lexicon as the reference translation. Natural languages, however, are characterized by their extremely rich mechanisms for reproduction via a large number of syntactic, lexical and semantic rewriting rules. Although BLEU has been shown to correlate positively with human assessments at the document level (Papineni et al., 2001), efforts to improve state-of-the-art MT require that human assessment be approximated at sentence level as well. Researchers report the BLEU score at document level in order to combat the sparseness of n-grams in BLEU scoring. But, ultimately, document-level MT evaluation has to be pinned down to the granularity of the sentence. Unfortunately, the correlation between human assessment and BLEU score at sentence level is extremely low (Liu et al., 2005, 2006). While acknowledging the appealing simplicity of BLEU as a way to access one perspective of an MT candidate translation's quality, we observe the following facts of n-gram based MT metrics. First, they may not reflect the mechanism of how human beings evaluate sentence translation quality. More specifically, optimizing BLEU does not guarantee the optimization of sentence quality approved by human assessors. Therefore, BLEU is likely to have a low correlation with human assessment at sentence level for most candidate translations. Second, it is conceivable that human beings are more reliable ranking the quality of multiple candidate translations than assigning a numeric value to index the quality of the candidate translation even with significant deliberation. Consequently, a more intuitive approach for automatic MT evaluation is to replicate the quality ranking ability of human assessors. Thirdly, the BLEU score is elusive and hard to interpret; for example, what can be concluded for a

Proceedings of the Second Workshop on Statistical Machine Translation, pages 240–247,
Prague, June 2007. ©2007 Association for Computational Linguistics

candidate translation's quality if the BLEU score is 0.0168, particularly when we are aware that even a human translation can receive an embarrassingly low BLEU score? In light of the discussion above, we propose an alternative scenario for MT evaluation, where, instead of assigning a numeric score to a candidate translation under evaluation, we predict its rank with regard to its peer candidate translations. This formulation of the MT evaluation task fills the gap between an automatic scoring function and human MT evaluation practice. The results from the current study will not only interest MT system evaluation moderators but will also inform the research community about which features are useful in improving the correlation between human rankings and automatic rankings.

2 Problem Formulation

2.1 Data and Human Annotation Reliability

We use two data sets for the experiments: the test data set from the LDC MTC corpus (LDC2003T17[1]) and the data set from the MT evaluation workshop at ACL05[2]. Both data sets are for Chinese-English language pairs and each has four reference translations and seven MT system translations as well as human assessments for fluency and adequacy on a scale of 1 to 5, with 5 indicating the best quality. For the LDC2003T17 data, human assessments exist for only three MT systems; for the ACL05 workshop data, there are human assessments for all seven MT systems. Table 1 summarizes the information from these two data sets.

The Kappa scores (Cohen, 1960) for the human assessment scores are negative, both for fluency and adequacy, indicating that human beings are not consistent when assigning quality scores to the candidate translations. We have much sympathy with a concern expressed in (Turian, 2003) that "Automatic MT evaluation cannot be faulted for poor correlation with the human judges, when the judges do not correlate well each other."To determine whether human assessor might be more consistent when ranking pairs of sentences, we examined the "ranking consistency score"of the human assessment data for the LDC2003T17 data. For this consistency score, we are only concerned with whether multiple judges are consistent in terms of which sentence of the two sentences is better: we are not concerned with the quantitative difference between judges. Since some sentences are judged by three judges while others are judged by only two judges, we calculated the consistency scores under both circumstances, referred to as "Consistent 2"and "Consistent 3"in the following table. For "Consistent 2", for every pair of sentences, where sentence 1 is scored higher (or lower or equal) than sentence 2 by both judges, then the two judges are deemed consistent. For "Consistent 3", the proportion of sentences that achieved the above consistency from triple judges is reported. Additionally, we also considered a consistency rate that excludes pairs for which only one judge says sentence 1 is better and the other judge(s) say(s) sentence 2 is better. We call these "Consistent 2 with tie"and "Consistent 3 with tie". From the rank consistency scores in Table 2, we observe that two annotators are more consistent with the relative rankings for sentence pairs than with the absolute quality scores. This finding further supports the task of ranking MT candidate sentences as more reliable than the one of classifying the quality labels.

2.2 Ranking Over Classification and Regression

As discussed in the previous section, it is difficult for human assessors to perform MT candidate translation evaluation with fine granularity (e.g., using real-valued numeric score). But humans' assessments are relatively reliable for judgments of quality ranking using a coarser ordinal scale, as we have seen above. Several approaches for automatically assigning quality scores to candidate sentences are available, including classification, regression or ranking, of which ranking is deemed to be a more appropriate approach. Nominalize the quality scores and formulating the task as a classification problem would result in a loss of the ordinal information encoded in the different scores. Additionally, the low Kappa scores in the human annotation reliability analysis reported above also confirms our previous speculation that a classification approach is less appropriate. Regression would be more reasonable than classification because it preserves the ordinal information in the quality labels, but it also inappropriately im-

[1] http://www.ldc.upenn.edu/Catalog/

[2] http://www.isi.edu/~ cyl/MTSE2005/

Data Index	MT Systems	References	Documents	Sentences
LDC2003T17	7	4	100	878
ACL05 Workshop	7	4	100	919

Table 1: Data Sets Information

Inter-Judge Score	Consistent 2	Consistent 3	Consistent 2 with Tie	Consistent 3 with Tie
Ranking Consistency Score	45.3%	23.4%	92.6%	87.0%

Table 2: Ranking Consisteny Scores for LDC2003T17 Data

poses interval scaling onto the quality labels. In contrast, ranking considers only the relative ranking information from human labels and does not impose any extra information onto the quality labels assigned by human beings.

The specific research question addressed in this paper is three-fold: First, in addition to investigating the correlation between automatic numeric scoring and human assessments, is ranking of peer candidate translations an alternative way of examining correlation that better suits the state of affairs of human annotation? Second, if the answer to the above question is yes, can better correlation be achieved with human assessment under the new task scenario? Finally, in addition to n-gram matching, which other knowledge sources can combat and even improve the rank order correlation? The process of ranking is a crucial technique in Information Retrieval (IR) where search engines rank web pages depending on their relevance to a query. In this work, sentence level MT evaluation is considered as a ranking problem. For all candidate translations of the same source Chinese sentence, we predict their translation quality ranks. We evaluate the ranker by Spearman's rank order correlation coefficient between human ranks and predicted ranks described by the following formula (Siegel,1956):

$$r = 1 - (\frac{6\sum D^2}{N(N^2-1)}) \qquad (1)$$

where D is the difference between each pair of ranks and N is the number of candidates for ranking.

3 Related Works

Papineni et al.(2001) pioneered the automatic MT evaluation study, which scores translation quality via n-gram matching between the candidate and reference translations. Following the growing awareness of the deficiency of n-gram based automatic MT evaluation, many studies attempted to improve upon n-gram based metrics (Zhou et al., 2006; Liu, et al., 2005,2006) as well as propose ways to evaluate MT evaluation metrics (Lin, et al. 2004). Previous studies, however, have focused on MT evaluation at the document level in order to fight n-gram sparseness problem. While document level correlation provides us with a general impression of the quality of an MT system, researchers desire to get more informative diagnostic evaluation at sentence level to improve the MT system instead of just an overall score that does not provide details. Recent years have seen several studies investigating MT evaluation at the sentence level (Liu et al., 2005,2006; Quirk, 2004). The state-of-the-art sentence level correlations reported in previous work between human assessments and automatic scoring are around 0.20. Kulesza et al.(2004) applied Support Vector Machine classification learning to sentence level MT evaluation and reported improved correlation with human judgment over BLEU. However, the classification taxonomy in their work is binary, being either machine translation or human translation. Additionally, as discussed above, the inconsistency from the human annotators weakens the legitimacy of the classification approach. Gamon et al.(2005) reported a study of English to French sentence-level MT evaluation without reference translations. In order to improve on the correlation between human assessments and the perplexity score alone, they combined a perplexity score with a classification score obtained from an SVM binary classifier distinguishing machine-translated sentences from human translations. The results showed that even the combination of the above two scores cannot outperform BLEU.

To sum up, very little consideration has been taken in previous research as to which learning approach is better motivated and justified by the state of affairs of human annotation reliability. Presumably, research that endeavors to emulate human performance on tasks that demontrate good inter-judge reliability is most useful.

a learning approach that is better supported by human annotation reliability can alleviate the noise from human assessments and therefore achieve more reliable correlations.

4 Experiments and Evaluation

4.1 Ranking SVM Learning Algorithm

Ranking peer candidate sentence translations is a task in which the translation instances are classified into a number of ranks. This is a canonical ordinal regression scenario, which differs from standard classification and metric regression. For implementation, we use the Ranking SVM of SVMlight (Joachims, 2004), which was originally developed to rank the web pages returned upon a certain query in search engines. Given an instance of a candidate translation, Ranking SVM assigns it a score based on:

$$U(x) = W^{\mathrm{T}}x \qquad (2)$$

where W represents a vector of weights (Xu et al., 2005). The higher the value of U(x), the better x is as a candidate translation. In an ordinal regression, the values of U(x) are mapped into intervals corresponding to the ordinal categories. An instance falling into one interval is classified into the corresponding translation quality. In ranking experiments, we use the Ranking SVM scores to rank the candidate sentences under evaluation.

4.2 Features

We experiment with three different knowledge sources in our ranking experiments:

1. N-gram matching between the candidate translation and the reference translation, for which we use BLEU scores calculated by the NIST script with smoothing[3] to avoid undefined log probabilities for zero n-gram probabilities.

2. Dependency relation matching between the candidate translation and the reference translation.

3. The log of the perplexity score of the candidate translation, where the perplexity score is obtained from a local language model trained on all sentences in the four reference translations using CMU SLM toolkit. The n-gram order is the default trigram.

4.2.1 N-gram matching feature

N-gram matching is certainly an important criterion in some cases for evaluating the translation quality of a candidate translation. We use the BLEU score calculated by the BLEU score script from NIST for this feature.

As has been observed by many researchers, BLEU fails to capture any non n-gram based matching between the reference and candidate translations. We carried out a pair-wise experiment on four reference translations from the LDC2003T17 test data, where we took one reference sentence as the reference and the other three references as candidate translations. Presumably, since the candidate sentences are near-optimal translations, the BLEU scores obtained in such a way should be close to 1. But our analysis shows a mean BLEU of only 0.1456398, with a standard deviation of 0.1522381, which means that BLEU is not very predictive of sentence level evaluation. The BLEU score is, however, still informative in judging the average MT system's translation.

4.2.2 Dependency Structure Matching

Dependency relation information has been widely used in Machine Translation in recent years. Fox (2002) reported that dependency trees correspond better across translation pairs than constituent trees. The information summarization community has also seen successful implementation of ideas similar to the depedency structure. Zhou et al.(2005) and Hovy et al.(2005) reported using Basic Elements (BE) in text summarization and its evaluation. In the current

[3] We added an extremely small number to both matched n-grams and total number of n-grams.

paper, we match a candidate translation with a reference translation on the following five dependency structure (DS) types:

- Agent - Verb
- Verb - Patient
- Modified Noun - Modifier
- Modified Verb - Modifier
- Preposition - Object

Besides the consideration of the presence of certain lexical items, DS captures information as to how the lexical items are assembled into a good sentence. By using their dependency relation match for ranking the quality of peer translations, we assume that the dependency structure in the source language should be well preserved in the target language and that multiple translations of the same source sentence should significantly share dependency structures. Liu et al.(2005) make use of dependency structure in sentence level machine translation evaluation in the form of headword chains, which are lexicalized dependency relations. We propose that unlexicalized dependency relations can also be informative. Previous research has shown that key dependency relations tend to have a strong correspondence between Chinese and English (Zhou et al., 2001). More than 80 % of subject-verb, adjective-noun and adverb-verb dependency relations were able to be mapped, although verb-object DS mapping is weaker at a rate of 64.8%. In our paper, we considered three levels of matching for dependency relation triplets, where a triplet consists of the DS type and the two lexical items as the arguments.

We used an in-house dependency parser to extract the dependency relations from the sentences. Figure 1 illustrates how dependency relation matching can go beyond n-gram matching. We calculated 15 DS scores for each sentence correponding to the counts of match for the 5 DS types at the 3 different levels.

4.2.3 Reference language model (RLM) feature

Statistical Language Modeling (SLM) is a key component in Statistical Machine Translation. The most dominant technology in SLM is n-gram models, which are typically trained on a large corpus for applications such as SMT and speech recognition. Depending on the size of the corpora used to train the language model, a language model can

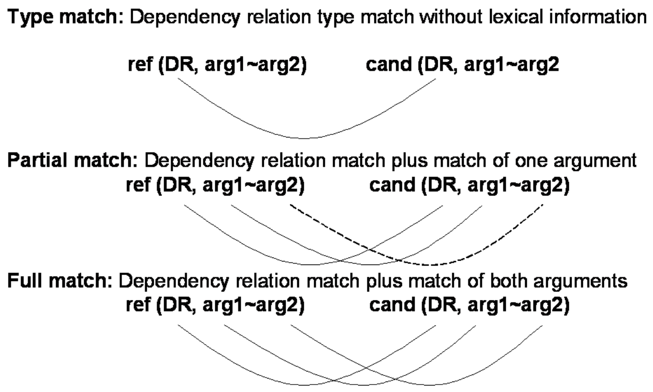
Type match: Dependency relation type match without lexical information
ref (DR, arg1~arg2)
cand (DR, arg1~arg2
Partial match: Dependency relation match plus match of one argument
ref (DR, arg1~arg2)
cand (DR, arg1~arg2)
Full match: Dependency relation match plus match of both arguments
ref (DR, arg1~arg2)
cand (DR, arg1~arg2)

Figure 1: Dependency Relation Matching Scheme

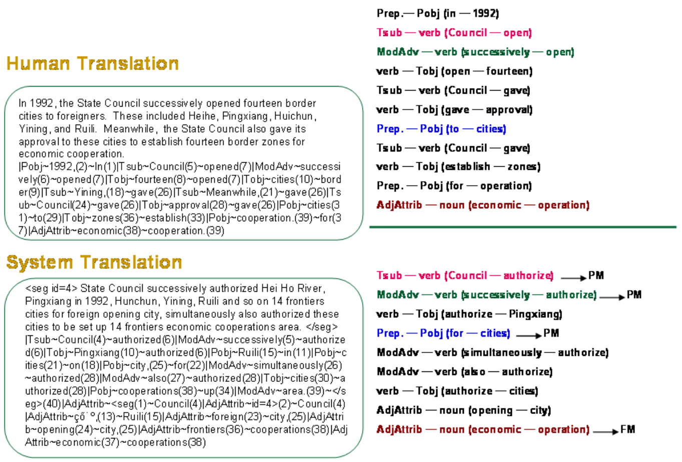
Human Translation
In 1992, the State Council successively opened fourteen border cities to foreigners. These included Heihe, Pingxiang, Huichun, Yining, and Ruili. Meanwhile, the State Council also gave its approval to these cities to establish fourteen border zones for economic cooperation.
|Pobj~1992,(2)~In(1)|Tsub~Council(5)~opened(7)|ModAdv~successively(6)~opened(7)|Tobj~fourteen(8)~opened(7)|Tobj~cities(10)~border(9)|Tsub~Yining,(18)~gave(26)|Tsub~Meanwhile,(21)~gave(26)|Tsub~Council(24)~gave(26)|Tobj~approval(28)~gave(26)|Pobj~cities(31)~to(29)|Tobj~zones(36)~establish(33)|Pobj~cooperation.(39)~for(37)|AdjAttrib~economic(38)~cooperation.(39)
Prep.— Pobj (in — 1992)
Tsub — verb (Council — open)
ModAdv — verb (successively — open)
verb — Tobj (open — fourteen)
Tsub — verb (Council — gave)
verb — Tobj (gave — approval)
Prep. — Pobj (to — cities)
Tsub — verb (Council — gave)
verb — Tobj (establish — zones)
Prep. — Pobj (for — operation)
AdjAttrib — noun (economic — operation)
System Translation
<seg id=4> State Council successively authorized Hei Ho River, Pingxiang in 1992, Hunchun, Yining, Ruili and so on 14 frontiers cities for foreign opening city, simultaneously also authorized these cities to be set up 14 frontiers economic cooperations area. </seg>
|Tsub~Council(4)~authorized(6)|ModAdv~successively(5)~authorized(6)|Tobj~Pingxiang(10)~authorized(6)|Pobj~Ruili(15)~in(11)|Pobj~cities(21)~on(18)|Pobj~city,(25)~for(22)|ModAdv~simultaneously(26)~authorized(28)|ModAdv~also(27)~authorized(28)|Tobj~cities(30)~authorized(28)|Pobj~cooperations(38)~up(34)|ModAdv~area.(39)~</seg>(40)|AdjAttrib~<seg(1)~Council(4)|AdjAttrib~id=4>(2)~Council(4)|AdjAttrib~çő˙°,(13)~Ruili(15)|AdjAttrib~foreign(23)~city,(25)|AdjAttrib~opening(24)~city,(25)|AdjAttrib~frontiers(36)~cooperations(38)|AdjAttrib~economic(37)~cooperations(38)
Tsub — verb (Council — authorize) → PM
ModAdv — verb (successively — authorize) → PM
verb — Tobj (authorize — Pingxiang)
Prep. — Pobj (for — cities) → PM
ModAdv — verb (simultaneously — authorize)
ModAdv — verb (also — authorize)
verb — Tobj (authorize — cities)
AdjAttrib — noun (opening — city)
AdjAttrib — noun (economic — operation) → FM

Figure 2: An Example - A Sentence Gets Credits for Dependency Relation Matching

be tuned to reflect n-gram probabilities for both a narrowed scope as well as a general scope covering the distribution of n-gram probabilities of the whole language. In the BLEU calculation, the candidate sentence is evaluated against an extremely local language model of merely the reference sentence. We speculate that a language model that stands in between such an immediate local language model and the large general English language model could help capture the variation of lexical and even structural selections in the translations by using information beyond the scope of the local sentence. Additionally, this language model could represent the style of a certain group of translators in a certain domain on the genre of news articles. To pursue such a language model, we explore a language model that is trained on all sentences in the four references. We obtain the perplexity score of each candidate sentence based on the reference language model. The perplexity score obtained this way reflects the degree to which a candidate translation can be generated from the n-gram probability distribution of the whole collection of sentences in the four references. It adds new information to BLEU because it not only compares the candidate sentence to its corresponding reference sentence but also reaches out to other sentences in the current document and other documents on the same topics. We choose perplexity over the language model score because the perplexity score is normalized with regard to the length of the sentence; that is, it does not favor sentences of relatively shorter length.

In our ranking experiments, for training, both the seven MT translations and the four reference translations of the same source sentence are evaluated as "candidate" translations, and then each of these eleven sentences is evaluated against the four reference sentences in turn. The BLEU score of each of these sentences is calculated with multiple references. Each dependency score is the average score of the four references. For the reference language model feature, the perplexity score is used for each sentence.

Conceptually, the reference language model and dependency structure features are more relevant to the fluency of the sentence than to the adequacy. Because the candidate sentences' adequacy scores are based on arbitrary reference sentences out of the four references in the human assessment data, we decided to focus on fluency ranking for this paper. The ranking scenario and features can easily be generalized to adequacy evaluation: the full and partial match dependency structure features are relevant to adeqaucy too. The high correlation between adequacy and fluency scores from human assessments (both pearson and spearman correlations are 0.67) also indicates that the same features will achieve improvements for adequacy evaluation.

Feature Set	Mean Corr	Corr Var
BLEU	0.3590644	0.0076498
DS	0.4002753	0.0061299
PERP	0.4273000	0.0014043
BLEU+DS	0.4128991	0.0027576
BLEU+PERP	0.4288112	0.0013783
PERP+DS	0.4313611	0.0014594
All	0.4310457	0.0014494

Table 3: Training and Testing on Within-year Data (Test on 7 MT and 4 Human)

4.3 Sentence Ranking on Within-year Data

In the first experiment, we performed the ranking experiment on the ACL05 workshop data and test on the same data set. We did three-fold cross-validation on two different test scenarios. On the first scenario, we tested the ranking models on the seven MT system output sentences and the four human reference sentences. It is widely agreed upon among researchers that a good evalutation metric should rank reference translation as higher than machine translation (Lin et al., 2004). We include the four human reference sentences into the ranking to test the ranker's ability to discriminate optimal translations from poor ones. For the second scenario, we test the ranking models on only the seven MT system output sentences. Because the quality differences across the seven system translations are more subtle, we are particularly interested in the ranking quality on those sentences. Tables 3 and 4 summarize the results from both scenarios.

The experimental results in the above tables conveyed several important messages: in the ranking setup, for both the MT and human mixed output and MT only output scenarios, we have a significantly

Feature Set	Mean Corr	Corr Var
BLEU	0.2913541	0.0324386
DS	0.3058766	0.0226442
PERP	0.2921684	0.0210605
BLEU+DS	0.315106	0.0206144
BLEU+PERP	0.2954833	0.0211094
PERP+DS	0.3067157	0.0217037
All	0.305248	0.0218777

Table 4: Training and Testing on Within-year Data (Test on MT only)

Feature Set	Mean Corr	Corr Var
BLEU	0.3133257	0.1957059
DS	0.4896355	0.0727430
PERP	0.4582005	0.0542485
BLEU+DS	0.4907745	0.0678395
BLEU+PERP	0.4577449	0.0563994
PERP+DS	0.4709567	0.0549708
All	0.4707289	0.0565538

Table 5: Training and Testing on Across-year Data (test on 3 MT plus 1 human)

improved correlation between human scoring and automatic ranking at sentence level compared to the state-of-the-art sentence level correlation for fluency score of approximately 0.202 found previously (Liu et al., 2006). When the ranking task is performed on a mixture of MT sentences and human translations, dependency structure and reference language model perplexity scores sequentially improve on BLEU in increasing the correlation. When the ranking task is performed only on MT system output sentences, dependency structure still significantly outperforms BLEU in increasing the correlation, and the reference language model, even trained on a small number of sentences, demonstrates utility equal to that of BLEU. The dependency structure feature proves to have robust utility in informing fluency quality in both scenarios, even with noise from the dependency parser, likely because a dependency triplet with inaccurate arguments is still rewarded as a type match or partial match. Additionally, the feature is reward-based and not penalty-based. We only reward matches and do not penalize mismatches, such that the impact of the noise from the MT system and the dependency parser is weakened.

4.4 Sentence Ranking on Across-year Data

It is trivial to retrain the ranking model and test on a new year's data. But we speculate that a model trained from a different data set can have almost the same ranking power as a model trained on the same data set. Therefore, we conducted an experiment where we trained the ranking model on the ACL 2005 workshop data and test on the LDC2003T17 data. We do not need to retrain the ranking SVM model; we only need to retrain the reference language model on the multiple references from the new year's data to obtain the perplexity scores. Because LDC2003T17 has human assessments for only three MT systems, we test on the three system outputs plus a human translation chosen randomly from the four reference translations. The results in Table 5 show an encouraging rank order correlation with human assessments. Similar to training and testing on within-year data, both dependency structure and perplexity scores achieve higher correlation than the BLEU score. Combining BLEU and dependency structure achieves the best correlation.

4.5 Document Level Ranking Testing

Previously, most researchers working on MT evaluation studied the correlation between automatic metric and human assessment on the granularity of the document to mitigate n-gram sparseness. Presumably, good correlation at sentence level should lead to good correlation at document level but not vice versa. Table 6 reports the correlations using the model trained on the 2005 workshop data and tested on the 100 documents of the LDC 2003 data. Comparing these correlations with the correlations reported in the previous section, we see that using the same model, the document level rank order correlation is substantially higher than the sentence level correlation, with the dependency structure showing the highest utility.

5 Conclusion and Future Work

The current study proposes to formulate MT evaluation as a ranking problem. We believe that a reliable ranker can inform the improvement of BLEU for a better automatic scoring function. Ranking in-

Feature Set	Mean Corr	Corr Var
BLEU	0.543	0.0853
DS	0.685	0.0723
PERP	0.575	0.0778
BLEU+DS	0.639	0.0773
BLEU+PERP	0.567	0.0785
PERP+DS	0.597	0.0861
All	0.599	0.0849

Table 6: Document Level Ranking Testing Results

formation could also be integrated into tuning process to better inform the optimization of weights of the different factors for SMT models. Our ranking experiments show a better correlation with human assessments at sentence level for fluency score compared to the previous non-ranking scenario, even with BLEU as the baseline feature. On top of BLEU, both the dependency structure and reference language model have shown encouraging utility for different testing scenarios. Looking toward the future work, more features could be explored, e.g., a parsing-based score of each candidate sentence and better engineering for dependency triplet extraction. Additionally, the entire research community on MT evaluation would benefit from a systematic and detailed analysis of real data that can provide a quantitative breakdown of the proportions of different "operations" needed to rewrite one sentence to another. Such an effort will guide MT evaluation researchers to decide which features to focus on.

References

J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37-46, 1960.

G. Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. HLT, pages 128–132, 2002.

H. J. Fox, Phrasal Cohesion and Statistical Machine Translation. EMNLP, 2002.

M. Gamon, et al., Sentence-level MT Evaluation without Reference Translations: Beyond Language Modeling, Proceedings of EAMT, 2005.

T. Joachims, Making Large-scale Support Vector Machine Learning Practical, in B. Scholkopf, C. Burges, A. Smola. Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA, December, 1998.

A. Kulesza and S. M. Shieber, A Learning Approach to Improving Sentence-Level MT Evaluation, 10th International Conference on Theoretical and Methodological Issues in Machine Translation, 2004.

C. Lin, et al., ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. COLING, 2004.

C. Lin, et al., Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, ACL, 2004.

D. Liu, et al., Syntactic Features for Evaluation of Machine Translation, ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.

D. Liu, et al., Stochastic Iterative Alignment for Machine Translation Evaluation, COLING/ACL Poster Session, Sydney, 2006.

C. B. Quirk, Training a Sentence-Level Machine Translation Confidence Measure, In Proceedings of LREC, 2004.

E. Hovy, et al., Evaluating DUC 2005 using Basic Elements. Document Understanding Conference (DUC-2005), 2005.

K. Papineni, et al., BLEU: a Method for Automatic Evaluation of Machine Translation, IBM research division technical report, RC22176 (W0109-022), 2001.

S. Siegel and N.J. Catellan, Non-parametric Statistics for the Behavioral Sciences, McGraw-Hill, 2nd edition, 1988.

M. Snover, et al., A Study of Translation Error Rate with Targeted Human Annotation, LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, 2005.

J. Turian, et al., Evaluation of Machine Translation and its Evaluation, MT Summit IX, 2003.

J. Xu, et al., Ranking Definitions with Supervised Learning Method, WWW'05 industry track, 811-819, 2005.

L. Zhou, et al., A BE-based Multi-document Summarizer with Query Interpretation. Document Understanding Conference (DUC-2005), 2005.

L. Zhou, C. Lin, E-evaluating Machine Translation Results with Paraphrase Support, EMNLP, 2006.

M. Zhou, C. Huang, Approach to the Chinese dependency formalism for the tagging of corpus. Journal of Chinese Information Processing, 8(3): 35-52, 1994.