2010 22nd International Conference on Tools with Artificial Intelligence

Support Vector Methods for Sentence Level Machine Translation Evaluation

Antoine Veillard*†, Elvina Melissa*, Cassandra Theodora*, Daniel Racoceanu† and Stéphane Bressan*
*School of Computing
National University of Singapore, Singapore 117417
†IPAL CNRS
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
Email: veillard@comp.nus.edu.sg

***Abstract*—Recent work in the field of machine translation (MT) evaluation suggests that sentence level evaluation based on machine learning (ML) can outperform the standard metrics such as BLEU, ROUGE and METEOR. We conducted a comprehensive empirical study on support vector methods for ML-based MT evaluation involving multi-class support vector machines (SVM) and support vector regression (SVR) with different kernel functions. We empathize on a systematic comparison study of multiple feature models obtained with feature selection and feature extraction techniques. Besides finding the conditions yielding the best empirical results, our study supports several unobvious conclusions regarding qualitative and quantitative aspects of feature sets in MT evaluation.**

I. Introduction

A. Sentence Level MT Evaluation

Sentence level MT evaluation is the assessment of the quality of a machine translated sentence. It is usually done by comparing the candidate translation to a reference translation.

Human judgement is the reference benchmark for MT evaluation. However, the ongoing research efforts in MT have created the need for more economical and practical alternatives to the evaluation by humans.

B. Standard Metrics for MT Evaluation

The use of handcrafted metrics is the current standard practice for automatic MT evaluation. The most commonly used metrics can be divided into 3 families: the BLEU [1] (and NIST [2]) family based on lexical matching, the newer METEOR [3] family introducing semantics and the ROUGE [4] family originally applied to evaluate text summaries. MaxSim [5] and the translation edit rate (TER) [6] are 2 other metrics involved in this work.

C. Machine Learning Approaches to MT Evaluation

Recent works, however, suggested that ML can also be used to create evaluation methods that effectively approximate human judgement. Corston-Oliver *et al.* [7] used decision trees to distinguish between human-produced and machine-produced sentences. Kulesza et Shieber [8], Albrecht and Hwa [9] and Ye *et al.* [10] applied support vector methods to different evaluation related tasks. Recently, Padó *et al.* [11] obtained good performances by using a regression-based approach.

D. Our Contribution

Further extending previous work on ML-based MT evaluation, we conducted systematic experiments using support vector methods with a strong focus on the feature model. In addition to determining how the different configurations of features, kernels and support vector methods perform in this task, we formulate a few conclusions regarding qualitative and quantitative aspects of feature sets for MT evaluation.

Due to space limitations, only a synthetic overview of our methods and results can be presented. We refer the reader to our technical report [12] for further details.

II. Method

A. Feature Model

Various feature models were created from a large pool of 239 initial features. They broadly divide into lexical features based on strings and grammatical features based on part-of-speech (POS) tags, the latter class being essentially original compared to previous ML-based approaches. The initial pool mostly contains basic features such as n-gram precision/recall rates but also more elaborate features including the metrics presented in Section I-B. Feature selection and extraction techniques were used to create feature models with a lower dimension. A complete description of the initial features and the resulting feature models is available in the technical report [12].

1) Feature Selection: Feature selection is performed by taking a subset of the initial features. Two different strategies were experimented:

- A categorical selection of features by separating lexical features from grammatical features or basic features from elaborate features.
- A selection of features based on their correlation (Pearson) with the quality of a translation.

2) Feature Extraction: Several feature models were obtained by performing a principal component analysis (PCA) on the data. The profile of the eigenvalues of the covariance matrix indicated that about 10 dimensions are sufficient to closely approximate the original data.

1082-3409/10 $26.00 © 2010 IEEE
DOI 10.1109/ICTAI.2010.122

B. Machine Learning Algorithms

We experimented with 2 different learning approaches, both based on support vector methods:

- A classification approach using a multi-class C-SVMs [13].
- A regression approach using ϵ-SVRs [14].

2 kernels were used: the linear kernel and the radial basis function (RBF) kernel.

III. Empirical Study

A. Dataset

Our dataset referred as the "NIST" dataset comprises 11028 sentence pairs from Chinese-English translations coming from the "Multiple-Translation Chinese Part 4" dataset produced by the Linguistic Data Consortium[1]. Each pair comes with 2 separate quality scores: *fluency* measuring the grammatical well-formedness and *adequacy* evaluating the preservation of the meaning.

B. Results

Performance results have been computed for all the 144 possible choices of feature model, learning algorithm, kernel and quality score. Only the main conclusions can be presented in this paper. Detailed data supporting the conclusions is available in the technical report [12].

All results correspond to the 5-fold cross-validation average of the Pearson correlation between the scores given by our ML-based metrics and the reference scores. Optimal SVM/SVR parameters were adjusted by grid-search.

C. Conclusions

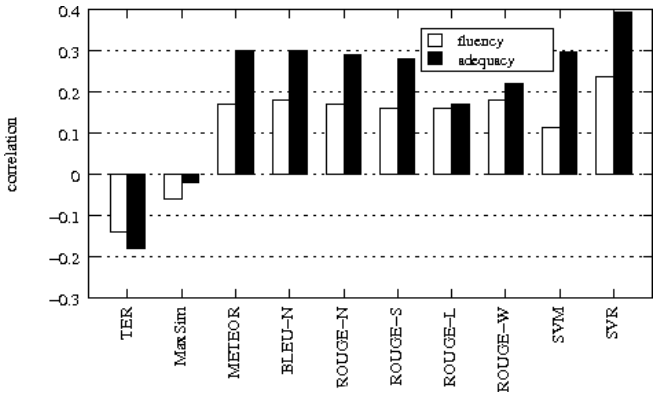

Fig. 1. Comparison of the best SVM and SVR results to maistream metrics. BLEU-N and ROUGE-N are given for the value of N yielding the best result.

- As illustrated by figure 1, ML-based metrics can largely outperform the most popular standard metrics in use for MT evaluation by as much as 34% for adequacy and 30% for fluency.
- Best results were achieved with an SVR using the RBF kernel on the full feature set. The SVR obtained significantly better results that the SVM, and the RBF kernel performed slightly better that the linear kernel.
- Feature sets based on POS tags alone largely outperformed standard metrics. This might be an indication that the importance of grammatical features is underestimated in the design of standard metrics.
- Sets of features showing poor individual Pearson correlation with quality scores (mainly high order n-gram precision/recall rates) yielded results almost on a par with the best results.
- Nearly optimal results obtained by the PCA-based feature sets prove that the data can be relevantly described by 10 or less dimensions, and that the 239 original features are therefore very redundant.

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

References

[1] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[2] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics," in *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT)*, San Diego, CA, USA, 2002, pp. 138–145.

[3] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL)*, Ann Arbor, MI, USA, 2005, pp. 65–72.

[4] C. Y. Lin, "Looking for a Few Good Metrics: ROUGE and its Evaluation," in *Proceedings of the 4th NTCIR Workshop*, Tokyo, Japan, 2004.

[5] Y. S. Chan and H. T. Ng, "MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, Colombus, OH, USA, 2008, pp. 55–62.

[6] M. Snover, B. J. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, MA, USA, 2006, pp. 223–231.

[7] S. Corston-Oliver, M. Gamon, and C. Brockett, "A Machine Learning Approach to the Automatic Evaluation of Machine Translation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, Toulouse, France, 2001, pp. 148–155.

[8] A. Kulesza and S. M. Shieber, "A Learning Approach to Improving Sentence-Level MT Evaluation," in *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Baltimore, MD, USA, 2004, pp. 75–84.

[9] J. S. Albrecht and R. Hwa, "A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, Prague, Czech Republic, 2007, pp. 880–887.

[10] Y. Ye, M. Zhou, and C.-Y. Lin, "Sentence Level Machine Translation Evaluation as a Ranking Problem: one Step Aside from BLEU," in *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT)*, Prague, Czech Republic, 2007, pp. 240–247.

[11] S. Padó, M. Galley, D. Jurafsky, and C. Manning, "Robust Machine Translation Evaluation with Entailment Features," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Singapore, 2009, pp. 297–305.

[12] A. Veillard, E. Melissa, C. Theodora, and S. Bressan, "TRC8/10: Support Vector Methods for Sentence Level Machine Translation Evaluation," National University of Singapore, Tech. Rep., 2010.

[13] C. Cortes and V. N. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, September 1995.

[14] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.