DFKI System Combination with Sentence Ranking at ML4HMT-2011

Eleftherios Avramidis
German Research Center for Artificial Intelligence (DFKI)
Language Technology Group (LT)
Berlin, Germany
eleftherios.avramidis@dfki.de

Abstract

We present a pilot study on a Hybrid Machine Translation system that takes advantag e of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting into a selection mechanism able to learn and rank system outputs on the sentence level, based on their quality. For training, due to the lack of human annotations, word-level Levenshtein distance has been used as a quality indicator, whereas a rich set of sentence features was extracted and selected from the dataset. Three classification algorithms (Naïve Bayes, SVM and Linear Regression) were trained and tested on pairwise featured sentence comparisons. The approaches yielded high correlation with original rankings (tau = 0.52) and selected the best translation in 54% of the cases.

1 Introduction

Optimizing Machine Translation (MT) performance through Hybrid Machine Translation has been a long standing goal, given the possible benefit from combining systems of different theoretical backgrounds (Habash, 2003). So far, research has adopted several approaches to MT system combinations. A vast majority of them treat the participating MT systems as black boxes, aiming to combine them based on some universal measure of quality (Callison-Burch and Flournoy, 2001). This has also allowed combinations of different outputs to take place on a word or phrase level (Matusov et al., 2006; Rosti et al., 2007; Hillard et al., 2007).

Meanwhile, there have been many suggestions that information derived from the translation process can contain useful hints for the quality of the produced output. Positive results have been shown on the development of Confidence Estimation metrics, in most of the cases complementing other universal features (Quirk, 2004; Rosti et al., 2007; Specia et al., 2009). Though, the best way to take advantage of such information, deriving from systems of different origin, remains still an open question.

Here we demonstrate a pilot study which tries to take advantage of the multi-dimensional and heterogeneous annotation annotation over MT output, provided in the frame of the ML4HMT-2011 shared task. In Section 2, we try to re-formulate the problem in a way which is easier to approach using Machine Learning (ML). In Section 3, we show how a suitable feature set has been extracted. In Section 4 we show the performance of Machine Learning algorithms and in Section 5 we provide a discussion of the results.

2 Re-formulation of the problem

2.1 Focus on a sentence level

The ML4HMT-2011 corpus provides a development and a test set of approximately 1,000 sentences each, translated by 5 different systems. Each translation output is accompanied with metadata referring to parts of the process each system performed. Although the annotation is rich, the main difficulty of the task relies on the fact that each system provides a different set of metadata, which are scattered over different derivation steps, that are not comparable through with other. For example, statistical systems provide statistics on the decoding steps and their search algorithm, while rule-based systems yield several derivation

steps within their tree analyses. For this reason, a simplified approach would be to restrict the granularity of the combination on the sentence level. This allows for a better picture on the compilation of the feature vectors that are required in a ML approach. It could also be applied for selecting the backbone translation in other MT combination approaches.

2.2 Pairwise decisions and ranking

Working on a sentence level leads to the goal of building an empirical selection mechanism, which would be able to estimate the quality of the generated sentence alternatives on the fly and choose accordingly. A draft learning approach on this direction would use a classification method, where the id of the best system serves as the class, and meta-data from all alternative outputs forms the feature vector for the classification. This approach, however, would result in a really difficult problem to solve, given also the size of the data, which would probably lead into sparseness problems.

Instead, we consider the tactic of breaking the quality judgement into pairwise comparisons, between all the 5 translation outputs per source sentence. This gives a total of about 17,000 training instances with binary classes, which makes the training of a classifier more plausible. Additionally, the classifier now has to "learn" and provide a binary answer to the much simpler question "*which of these two sentences is better?*", given the meta-data from the two systems themselves. The pairwise (positive or negative) judgments are then summed up, so as to order the 5 outputs based on their predicted quality. We have therefore reformulated the problem into modelling a *quality ranking* of the sentences. Coming back to the system combination requirement, the best ranked sentence can then be selected for the combined output.

2.3 Supervised learning

It would make sense to try to learn such a mechanism, given a training set with relatively reliable quality indicators, for example, results of human evaluation. Unfortunately, although a development set has been provided, it does not include an objective measurement of quality within each set of 5 alternatives. The only relevant information can be derived from the reference translation, which, in a way, forms the gold translation that that the MT systems should reach.

As an answer to this question, we examined the so-called segment-level metrics that could provide this information. In the end, word-level Levenshtein distance seemed to adhere better to our needs. So, thereafter we consider this as a quality indication and we will develop and evaluate the ML outcome based on it. This would provide us with an intuition for the learning capabilities of the approach and allow a potential shift to gold human judgments, when these are available.

3 Extracting and selecting features

Given the decisions described above, the various multilateral and overlapping annotations on several levels of the translation process have to be converted to a shallow set of sentence-level features.

3.1 Defining sentence-level features

Based on our intuition given the knowledge about the functioning of each system, we extracted the following features:

- **Joshua**: overall translation probability, tuned weights, count of phrases. Decoding search features included the number of pre-pruned, added, merged nodes, and of fuzzy matches. Three sentence-level statistics were derived from the sequence of feature scores for every decoding step leading to the dominant output: average, standard deviation and variance.

- **MaTrEx**: overall translation probability, tuned weights, count of phrases. As done above, three sentence-level statistics were derived from the sequence of phrase scores and future cost estimates: average, standard deviation and variance.

- **Lucy**: indication that the system performed phrasal analysis and segment combination in the transfer phase (Federmann and Hunsicker, 2011), counts of all nodes appearing in the derivation trees.

- **All**: Scores provided by external linguistic analysis tools, including language model probability (bi-gram, tri-gram, 5-gram), PCFG parsing score (ratio of target to source), number of tokens, number of unknown words. This information was needed for the systems which had no other features easily extractable.

feature	Inf. gain	Gain ratio	Gini
Lucy phrasal analysis	0.181	0.092	0.059
Joshua total probability	0.100	0.050	0.030
External 5gram score	0.000	0.037	0.000
MaTrEx std deviation of future cost	0.058	0.029	0.019
MaTrEx std deviation of probabilities	0.058	0.029	0.019
Joshua/MaTrEx phrase count	0.012	0.005	0.004

Table 1: Results of feature selection by Information Gain, Gain Ratio and Gini Index

feature	ReliefF
Joshua total probability	0.064
Lucy phrasal analysis	0.023
MaTrEx total probability	0.012
Joshua merged nodes	0.011
Joshua word penalty variance	0.010

Table 2: Results of ReliefF feature selection

classifier	p.ac.	τ	b.ac
SVM	0.52	0.52	0.53
Bayes	0.63	0.43	0.54
Linear	0.51	0.25	0.50

Table 3: Results of the classification process

N-gram features have been generated with the SRILM toolkit (Stolcke, 2002) using a language model trained over all monolingual training sets for the WMT 2011 Shared Task (Callison-Burch et al., 2011), interpolated on the 2007 test set. PCFG parsing was done with the Berkeley Parser (Petrov and Klein, 2007), trained over an English and a Spanish treebank (Mariona Taulé and Recasens, 2008). The feature selection algorithms (as well as the learning algorithms below) were implemented with the Orange toolkit (Demšar et al., 2004).

3.2 Feature selection

The whole extraction process, despite the fact that many other annotations were ignored, resulted in a set of more than 50 features per sentence (particularly due to the counts of tree tags). Many machine learning algorithms perform better when they are provided rather smaller sets of uncorrelated features. Even for the algorithms that perform sentence selection themselves, big sets increase the complexity and required runtime.

Three feature selection algorithms were examined as a first step. We computed scores for all attributes based on ReliefF (Kononenko, 1994), Information Gain (Kullback and Leibler, 1951), Gain Ratio and Gini index (Ceriani and Verme, 2011), which can be seen in Tables 1 and 2. We chose the features that have a score higher than 0.01 in either of the metrics.

4 Machine learning algorithms

For the actual task of learning the pairwise comparisons, we trained a SVM, a Naïve Bayes (Cleveland, 1979) and a linear classifier. Feature selection was applied for the latter two, as well as imputation for the missing feature values. Due to implementation issues SVM was lacking the features of category "all" (Section 3). We computed the *pairwise accuracy* (p.ac.) of the classification, the *segment-level tau coefficient* (τ), which indicates the correlation with the rankings produced with word-level Levenshtein distance and the accuracy when focusing only on whether the *best rank* was predicted (b.ac), all measured over the test set. The results can be seen in table 3

5 Discussion

Best-rank accuracy indicates that the classifiers managed to provide the best solution right away, in 50-54% of the cases. This is relatively low, but it can be still considered a small success, given the fact that the probability of random selection out of the five alternatives would be 20%. With some manual evaluation look-up in the classification performed by SVM, we were able to draw the conclusion that this has mostly to do with the fact that the classifier comes to a level of uncertainty concerning the two best ranked sentences. So,

most of the times, contradictory judgments would lead to a tie for the two best scored systems, although only one of them needs to be selected. We believe that further processing needs to take place, so that ties as a result of uncertain classification, particularly for the first rank, can be eliminated.

The classifier built with SVM gives the best average sentence-level correlation. This means that it predicted the ranking of the systems better than the other systems, although there were mistakes. Though, the reproduced ranking was rarely too bad, since only 6% of the sentences had a negative tau coefficient. We can also note that the tau correlation given in this task is much higher than the ones achieved by evaluation metrics in WMT Shared Tasks (Callison-Burch et al., 2011), which go up to $\tau = 0.35$. Though, human rankings are not comparable with Levenshtein distance rankings, therefore no clear comparison can be done.

6 Conclusion

We presented an effort to reduce Hybrid Machine Translation selection into sentence-level ranking. Features extracted from the sentence level have been used to train three classification algorithms. SVM shows high sentence-level correlation with the original quality score, whereas Naïve Bayes succeeds slightly better into choosing the best translation per sentence. The potential for further improvement, with more sophisticated feature extraction should be examined.

Acknowledgments

This work has been developed within the TaraXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. Many thanks to Lukas Poustka for preprocessing and training the Spanish grammar.

References

Callison-Burch, C. and Flournoy, R. S. (2001). A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of the Machine Translation Summit VIII*, pages 63–66.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Ceriani, L. and Verme, P. (2011). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, pages 1–23. 10.1007/s10888-011-9188-x.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Demšar, J., Zupan, B., Leban, G., and Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. In *Principles of Data Mining and Knowledge Discovery*, pages 537–539.

Federmann, C. and Hunsicker, S. (2011). Stochastic parse tree selection for an existing rbmt system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 351–357, Edinburgh, Scotland. Association for Computational Linguistics.

Habash, N. Y. (2003). *Generation-heavy hybrid machine translation*. PhD thesis, University of Maryland at College Park, College Park, MD, USA. AAI3094491.

Hillard, D., Hoffmeister, B., Ostendorf, M., Schlueter, R., and Ney, H. (2007). iROVER: Improving system combination with classification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 65–68, Rochester, New York. Association for Computational Linguistics.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *Proceedings of the European conference on machine learning on Machine Learning*, pages 171–182, Secaucus, NJ, USA. Springer-Verlag New York, Inc.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

Mariona Taulé, M. A. M. and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*,

Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Matusov, E., Ueffing, N., and Ney, H. (2006). *Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment*, pages 33–40. Association for Computational Linguistics.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Quirk, C. B. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Protugal. European Language Resources Association (ELRA).

Rosti, A.-V., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. (2007). Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York. Association for Computational Linguistics.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M., and Cristianini, N. (2009). Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the EAMT*, pages 28–35. European Association for Machine Translation.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904, Denver, Colorado, USA. ISCA Archive.