

Results from the ML4HMT-12 Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation

*Christian FEDERMANN*¹ *Tsuyoshi OKITA*² *Maite MELERO*³
*Marta R. COSTA—JUSSÀ*³ *Toni BADIA*³ *Josef VAN GENABITH*²

(1) DFKI GmbH, Language Technology Lab, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, GERMANY

(2) Dublin City University, School of Computing, Glasnevin, Dublin 9, IRELAND

(3) Barcelona Media, Speech and Language Group, Diagonal 177, 08018 Barcelona, SPAIN

`cfedermann@dfki.de {tokita, josef}@computing.dcu.ie`

`{maite.melero, marta.ruiz, toni.badia}@barcelonamedia.org`

Abstract

We describe the second edition of the ML4HMT shared task which challenges participants to create hybrid translations from the translation output of several individual MT systems. We provide an overview of the shared task and the data made available to participants before briefly describing the individual systems. We report on the results using automatic evaluation metrics and conclude with a summary of ML4HMT-12 and an outlook to future work.

Keywords: Machine Translation, System Combination, Machine Learning.

1 Introduction

The ML4HMT-12 workshop and associated shared task are an effort to trigger a systematic investigation on improving state-of-the-art hybrid machine translation, making use of advanced machine-learning (ML) methodologies. The first edition of the workshop (ML4HMT-11) also road-tested a shared task (and associated data set) described and summarised in (Federmann, 2011). The main focus of the ML4HMT-12 (and ML4HMT-11) shared task is to address the question:

Can Hybrid MT and System Combination techniques benefit from extra information (linguistically motivated, decoding, runtime, confidence scores or other meta-data) from the individual MT systems involved?

Participants are invited to build hybrid MT systems and/or system combinations by using the output of several MT systems of different types, as provided by the organisers. While participants are encouraged to explore machine learning techniques to explore the additional meta-data information sources, other approaches aimed at general improvements in hybrid and combination based MT are welcome to participate in the challenge. For systems that exploit additional meta-data information the challenge is that additional meta-data is highly heterogeneous and specific to individual systems.

One of the core objectives of the challenge is to build an MT combination (or more generally a hybrid MT) mechanism, where possible making effective use of the system-specific MT meta-data output produced by the participating individual MT systems as provided by the challenge development set data comprising outputs of four distinct MT systems and various meta-data annotations. The development set provided by the organisers can be used for tuning the combination or hybrid systems during the development phase.

2 Datasets

The organisers of the ML4HMT-12 shared task provide two data sets, one for the language pair Spanish→English (ES-EN), the other for Chinese→English (ZH-EN).

ES-EN Participants are given a development bilingual data set aligned at a sentence level. Each "bilingual sentence" contains:

1. the source sentence;
2. the target (reference) sentence; and
3. the corresponding translations from four individual component MT systems, based on different machine translation paradigms (Apertium (Ramírez-Sánchez et al., 2006); Lucy (Alonso and Thurmair, 2003); two different variants of Moses (Koehn et al., 2007): PB-SMT and HPB-SMT).

The output has been automatically annotated with system-internal meta-data information derived from the translation process of each of the systems.

ZH-EN A corresponding data set for Chinese→English with output translations from three systems (Moses; ICT_Chiero (Mi et al., 2009); Huajian RBMT) was prepared. Again, system output has been automatically annotated with system-internal meta-data information.

In total, with the development data participants received 20,000 translations per system for training and had to translate a test set containing 3,003 sentences ("newstest2011") for Spanish→English, while for the other language pair Chinese→English, a total of 6,752 training sentences per system were available while the test set had a size of 1,357 sentences.

3 Participants

We received six submissions for the Spanish→English translation task and none for Chinese→English. Below, we will briefly describe the participating systems.

3.1 DCU-Alignment

The authors of (Wu et al., 2012) incorporate alignment information as additional meta-data into their system combination module which does not originally utilise any alignment information provided by the individual MT systems producing the candidate translations. The authors add alignment information provided by one of the MT systems, the Lucy RBMT engine, into the internal, monolingual, alignment process. Unfortunately, the extracted alignment is often already a subset of alignments calculated by the monolingual aligner in the system combination and hence the approach does not augment the overall system combination performance as much as expected.

3.2 DCU-QE1

The submission described in (Okita et al., 2012a) incorporates a sentence-level Quality Estimation (QE) score as meta-data into their system combination module. Recently, QE or confidence estimation technology has advanced. It measures the quality of translations without references. The core idea is to incorporate this knowledge into the system combination module through an improved backbone selection.

3.3 DCU-QE2

The work described in (Okita et al., 2012a) also explains how one can incorporate a sentence-level Quality Estimation score to do the data selection process. The authors designed, hence, a method only based on Machine Learning. The translated output tends to preserve the translation quality as is expected, which results in a high Meteor score. The idea in this paper is to select one of the given translation outputs by QE score where a sentence-level QE technology is to measure the confidence estimation for the translation output.

3.4 DCU-DA

The authors of (Okita et al., 2012b) utilised unsupervised topic/genre classification results as meta-data, feeding into their system combination module. Since this module has access to topic/genre information, an MT system can take advantage of this information. MT systems are tuned to particular topic/genre groups and only compute translations for documents in this group, hence the performance of such MT systems may improve.

3.5 DCU-LM

This submission incorporates latent variables as meta-data into the system combination module. Information about those latent variables are supplied by a probabilistic neural language model. This language model can be trained on a huge monolingual corpus, with the disadvantage that the training of such a model takes considerable time. In fact, the LM used for ML4HMT-12 was small due to the huge cost of training, resulting in only small performance gains.

| Score | Spanish→English | | | | | |
|--------|-----------------|---------|---------------|---------|---------|----------------|
| | 1best | R3 | DCU-DA | DCU-LM | DCU-QE1 | DFKI |
| Meteor | 0.30692 | 0.32226 | 0.32124 | 0.31684 | 0.31712 | 0.32303 |
| NIST | 7.4296 | 7.4291 | 7.6771 | 7.5642 | 7.6481 | 7.2830 |
| BLEU | 0.2614 | 0.2524 | 0.2634 | 0.2562 | 0.2587 | 0.2570 |

Table 1: Translation quality of ML4HMT-12 submissions measured using Meteor, NIST, and BLEU scores for language pair Spanish→English. Best system per metric printed in bold face.

3.6 DFKI

This submission implements a method for system combination based on joint, binarised feature vectors as introduced in (Federmann, 2012b). It can be used to combine several black-box source systems. The authors first define a total order on the given translation output which can be used to partition an n -best list of translations into a set of pairwise system comparisons. Using this data, they train an SVM-based classification model and show how this classifier can be applied to combine translation output on the sentence level.

4 Results

Similar to the first edition of the ML4HMT shared task (ML4HMT-11), we aim to run both an automatic and a manual evaluation campaign. We consider three automatic scoring metrics, namely Meteor (Denkowski and Lavie, 2011), NIST (Dodgington, 2002), and BLEU (Papineni et al., 2002), which are all well-renowned evaluation metrics commonly used for MT evaluation. Manual evaluation is currently being conducted using the Appraise software toolkit as described in (Federmann, 2012a). Table 1 summarises the results for all participating systems.

5 Conclusion

System *DFKI* performed best in terms of Meteor score¹ while system *DCU-DA* achieved best performance for NIST and BLEU scores. It will be interesting to see how these findings correlate with the results from manual evaluation, something we will report on in future work.

If technically feasible, we also intend to apply the algorithms submitted to the Spanish→English portion of the Shared Task to the second language pair, Chinese→English.

Acknowledgments

This work has been supported by the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement: 249119) as well as by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (<http://www.cngl.ie/>) at Dublin City University. We are grateful to the organisers of COLING for supporting ML4HMT-12. Also, we are greatly indebted to all members of our program committee for their excellent and timely reviewing work and useful comments and feedback. Last but not least we want to express our sincere gratitude to all participants of the shared task for their interesting submissions.

¹Which, in the first edition of the ML4HMT shared task, had shown the best correlation with human judgments. A finding that will be investigated in more detail once results from the manual evaluation of ML4HMT-12 are available.

References

- Alonso, J. A. and Thurmair, G. (2003). The Compendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.
- Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. ACL.
- Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Federmann, C. (2011). Results from the ML4HMT Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4. Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-11))*, pages 110–117, Barcelona, Spain. META-NET.
- Federmann, C. (2012a). Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:25–35.
- Federmann, C. (2012b). A Machine-Learning Framework for Hybrid Machine Translation. In *Proceedings of the 35th Annual German Conference on Artificial Intelligence (KI-2012)*, pages 37–48, Saarbrücken, Germany. Springer, Heidelberg.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL Demo and Poster Sessions*, pages 177–180. ACL.
- Mi, H., Liu, Y., Xia, T., Xiao, X., Feng, Y., Xie, J., Xiong, H., Tu, Z., Zheng, D., Lu, Y., and Liu, Q. (2009). The ICT Statistical Machine Translation Systems for the IWSLT 2009. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 55–59, Tokyo, Japan.
- Okita, T., Rubino, R., and van Genabith, J. (2012a). Sentence-level quality estimation for mt system combination. In *Proceedings of the Second Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-12)*, Mumbai, India. Accepted for publication.
- Okita, T., Toral, A., and van Genabith, J. (2012b). Topic modeling-based domain adaptation for system combination. In *Proceedings of the Second Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-12)*, Mumbai, India. Accepted for publication.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA. ACL.

Ramírez-Sánchez, G., Sánchez-Martínez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A., and Forcada, M. L. (2006). Openrad apterium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*.

Wu, X., Okita, T., van Genabith, J., and Liu, Q. (2012). System combination with extra alignment information. In *Proceedings of the Second Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (MLAHMT-12)*, Mumbai, India. Accepted for publication.