

# The ML4HMT Workshop on Optimising the Division of Labour in Hybrid Machine Translation

Christian Federmann<sup>4</sup>, Eleftherios Avramidis<sup>4</sup>, Marta R. Costa-jussà<sup>1</sup>,  
Maite Melero<sup>1</sup>, Pavel Pecina<sup>2</sup>, Josef van Genabith<sup>3</sup>

<sup>1</sup> Barcelona Media, Barcelona, Spain

<sup>2</sup> Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

<sup>3</sup> DCU Dublin City University, Dublin, Ireland

<sup>4</sup> DFKI GmbH, Berlin & Saarbrücken, Germany

cfedermann@dfki.de, eleftherios.avramidis@dfki.de, marta.ruiz@barcelonamedia.org,  
maite.melero@barcelonamedia.org, pecina@ufal.mff.cuni.cz, josef@computing.dcu.ie

## Abstract

We describe the “Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation” (ML4HMT) which aims to foster research on improved system combination approaches for machine translation (MT). Participants of the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. We first describe the ML4HMT corpus used in the shared task, then explain the XLIFF-based annotation format we have designed for it, and briefly summarize the participating systems. Using both automated metrics scores and extensive manual evaluation, we discuss the individual performance of the various systems. An interesting result from the shared task is the fact that we were able to observe different systems winning according to the automated metrics scores when compared to the results from the manual evaluation. We conclude by summarising the first edition of the challenge and by giving an outlook to future work.

**Keywords:** Machine Translation, System Combination, Machine Learning

## 1. Introduction

The “Challenge on Optimising the Division of Labour in Hybrid Machine Translation” is an attempt to trigger systematic investigation on improvements of state-of-the-art hybrid machine translation (MT), using advanced machine-learning (ML) methodologies. Participants of the challenge are requested to design hybrid MT or system combination methods, combining the translation output of several systems of different types, which is provided by the organisers. The main focus of the shared task is trying to answer the following question:

*Can hybrid machine translation or system combination techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

Our research is part of the META-NET project and focuses on the design and development of such advanced combination approaches, possibly bridging the gap to the machine learning community to foster joint and systematic exploration of novel system combination techniques; for this, we have collected translation output from various machine translation systems, annotated their output with information such as part-of-speech, word alignment, or language model scores. The collected data has been released as a multilingual corpus. Furthermore, we have organised a workshop and a challenge exploiting the ML4HMT corpus.

In this paper, we describe the data given to the shared task participants (Section 2), present the systems taking part in the challenge (Section 3), and report results based on both automated metrics’ scores and manual evaluation efforts

(Section 4). We conclude by giving a summary and an outlook to future work in Section 5.

## 2. Data

The participants are given a bilingual development set, aligned at a sentence level. Each “bilingual sentence” contains:

- the source sentence;
- the target (reference) sentence;
- the corresponding output translations from five different systems, based on different MT approaches: Apertium (Ramírez-Sánchez et al., 2006), Joshua (Li et al., 2009), Lucy (Alonso and Thurmair, 2003), Matrex (Penkale et al., 2010), and Metis (Vandeghinste et al., 2008). The output has been annotated with system-internal information derived from the translation process of each of the systems.

### 2.1. Annotated Data Format

We have developed an annotation format derived from XLIFF (XML Localisation Interchange File Format) to represent and store the corpus data. XLIFF is an XML-based format created to standardize localisation. It was standardised by OASIS in 2002 and its current specification is v1.2 released on February 1, 2008 (<http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html>).

An XLIFF document is composed of one or more <file> elements, each corresponding to an original file or source. Each <file> element contains the source of the data to

be localised and the corresponding localised (translated) data for one locale only. The localisable texts are stored in `<trans-unit>` elements, each containing a `<source>` element to store the source text and a `<target>` (not mandatory) element to store the translation.

We introduced new elements into the basic XLIFF format (inside a new "metanet" namespace) allowing to add a wide variety of meta-data annotations of the translated texts by different MT systems (tools). The tool information is included in the `<tool>` element appearing in the header of the file. Each tool can have one or more parameters (model weights) which are described in the `<metanet:weight>`.

Annotation is stored in an `<alt-trans>` element within the `<trans-unit>` elements. The `<source>` and `<target>` elements inside `<trans-unit>` refer to the source sentence and its reference translation, respectively. The `<source>` and `<target>` elements in the `<alt-trans>` elements represent the input text and corresponding translation output of a particular MT system (tool). Tool-specific scores assigned to the translated sentence are listed in the `<metanet:scores>` element and the derivation of the translation is specified in the `<metanet:derivation>`.

## 2.2. Development and Test Sets

We decided to use the WMT 2008 (Callison-Burch et al., 2008) news test set as a source for the annotated corpus. This is a set of 2,051 sentences from the news domain, translated to several languages, including English and Spanish but also others. The data was provided by the organisers of the Third Workshop on Machine Translation (WMT) in 2008. This data set was split into our own development set (containing 1,025 sentence pairs) and test set (containing 1,026 sentence pairs).

## 3. Participating Systems

### 3.1. DCU

The authors of (Okita and van Genabith, 2011) describe a system combination module in the MaTrEx (Machine Translation using Examples) system developed at Dublin City University. The system combination module for the shared task achieved an improvement of 2.16 BLEU (Papineni et al., 2001) points absolute and 9.2% relative compared to the best single system, which did not use any external language resources. The DCU system uses a confusion network on top of a Minimum Bayes Risk decoder (MBR decoder) (Kumar and Byrne, 2002), which has recently become a popular technique (Bangalore et al., 2001; Matusov et al., 2006).

### 3.2. DFKI-A

The system described in (Avramidis, 2011) reported on translations from a system combination with a sentence ranking component. The proposed solution offers a machine learning approach, resulting in a selection mechanism able to learn and rank systems' translation output on the sentence level, based on their respective quality.

For training, due to the lack of human annotations, word-level Levenshtein distance (Levenshtein, 1966) has been

used as a (minimal) quality indicator, whereas a rich set of sentence features was extracted and selected from the dataset. Three classification algorithms (Naive 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 up to 54% of the cases.

### 3.3. DFKI-B

The authors of (Federmann et al., 2011) report on experiments that are based on factored word substitution. Out of the data provided by the workshop organisers, they choose one system to provide the "translation backbone". The other four systems are mined for alternate translations that are potentially substituted into the aforementioned template translation if the system finds enough evidence that the candidate translation is better. Each of these substitution candidates is evaluated concerning a number of factors: 1) part-of-speech, 2) language model scores, 3) context.

### 3.4. LIUM

Patrik Lambert (LIUM) has submitted results from applying the open-source MANY system (Barrault, 2010) on our data set. The MANY system can be decomposed in two main modules.

The first one is the alignment module which actually is a modified version of TERp (Snover et al., 2009). Its role is to incrementally align the hypotheses against a backbone in order to create a confusion network. Each hypothesis acts as backbone, yielding each the corresponding confusion network. Those confusion networks are then connected together to create a lattice.

The second module is the decoder which is based on the token pass algorithm and operating on the lattice created previously. Future costs can be computed as a weighted sum of the logarithm of feature functions.

## 4. Evaluation Results

To evaluate the performance of the participating systems, we computed automated scores, namely BLEU, NIST (Dodgington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann et al., 1997), Word error rate (WER) (Hunt, 1990), and Translation Error Rate (TER) (Snover et al., 2006) and also performed an extensive, manual evaluation with 3 annotators ranking system combination results for a total of 904 sentences.

### 4.1. Automated Scores

Results from running automated scoring tools on the submitted translations are reported in Table 1. The overall best value for each of the scoring metrics is print in **bold face**. Table 2 presents automated metric scores for the individual systems in the ML4HMT corpus, also computed on the test set. These scores give an indicative baseline for comparison with the system combination results.

### 4.2. Manual Ranking

The manual evaluation is undertaken using the Appraise (Federmann, 2010) system; a screenshot of the evaluation interface is shown in Figure 1. Users are shown a reference

sentence and the translation output from all four participating systems and have to decide on a ranking in *best-to-worst order*. Table 3 shows the average ranks per system from the manual evaluation, again the best value per column is printed in **bold face**. Table 4 gives the statistical mode per system which is the value that occurs most frequently in a data set.

### 4.3. Inter-annotator Agreement

Next to computing the average rank per system and the statistical mode, we follow (Bojar et al., 2011) and compute Scott’s  $\pi$  scores to be comparable to WMT11. In our manual evaluation campaign, we had  $n = 3$  annotators assign ranks to our four participating systems. As ties were not allowed, this means there exist  $4! = 24$  possible rankings per sentence (e.g., *ABCD*, *ABDC*, etc.). Overall, we collected rankings for 904 sentences with an overlap of  $N = 146$  sentences for which all annotators assigned ranks.

Scott’s  $\pi$  allows to measure the pairwise annotator agreement for a classification task. It is defined as

$$\pi = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

where  $P(A)$  represents the fraction of rankings on which the annotators agree, and  $P(E)$  is the probability that they agree by chance. Table 5 lists the pairwise agreement of annotators for all four participating systems. Assuming  $P(E) = 0.5$  we obtain an overall agreement  $\pi$  score of

$$\pi = \frac{0.673 - 0.5}{1 - 0.5} = 0.346 \quad (2)$$

which can be interpreted as *fair agreement* following Landis and Koch (1977). WMT shared tasks have shown this level of agreement is common for language pairs, where the performance of all systems is rather close to each other, which in our case is indicated by the small difference measured by automatic metrics on the test set (Table 1). The lack of ties, in this case might have meant an extra reason for disagreement, as annotators were forced to distinguish a quality difference which otherwise might have been annotated as “equal”.

## 5. Conclusion

We have developed an annotated hybrid sample MT corpus which contains a set of 2,051 sentences translated by five different MT systems<sup>1</sup> (Joshua, Lucy, Metis, Apertium, and MaTrEx). Using this resource we have launched the Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT (ML4HMT-2011), asking participants to create combined, hybrid translations using machine learning algorithms or other, novel ideas for making best use of the provided ML4HMT corpus data.

Four participating combination systems, each following a different solution strategy, have been submitted to the shared task. We computed automated metrics’ scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly,

the system winning nearly all the automatic scores only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings ranked last place in the automatic metric scores based evaluation. This clearly indicates that more systematic investigation of hybrid system combination approaches, both on a system level and regarding the evaluation of such systems’ output, needs to be undertaken. We have learned from the participants that some of the meta-data annotations contributed by the individual MT systems in our ML4HMT corpus are too heterogeneous to be used easily in system combination approaches; hence we will work on an updated version for the next edition of this shared task. Also, we will further focus on the integration of advanced machine learning techniques as these are expected to support better exploitation of our corpus’ data properties.

## Acknowledgments

This work has been funded under the Seventh Framework Programme of the European Commission through the T4ME contract (grant agreement no.: 249119) and was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation ([www.cngl.ie](http://www.cngl.ie)) at Dublin City University. Also, it was partially supported by the Czech Science Foundation (grant no. P103/12/G084) partially funded by the Spanish Ministry of Economy and Competivity through the Juan de la Cierva fellowship program. We are grateful to the anonymous reviewers for their valuable feedback.

## 6. References

- Juan A. Alonso and Gregor Thurmair. 2003. The Compendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.
- Eleftherios Avramidis. 2011. DFKI System Combination with Sentence Ranking at ML4HMT-2011. 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 (ML4HMT)*, Barcelona, Spain, November. META-NET.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *ASRU*, Italy.
- Loïc Barrault. 2010. MANY : Open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93:147–155.
- Ondrej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh,

<sup>1</sup>Not all systems available for all language pairs.

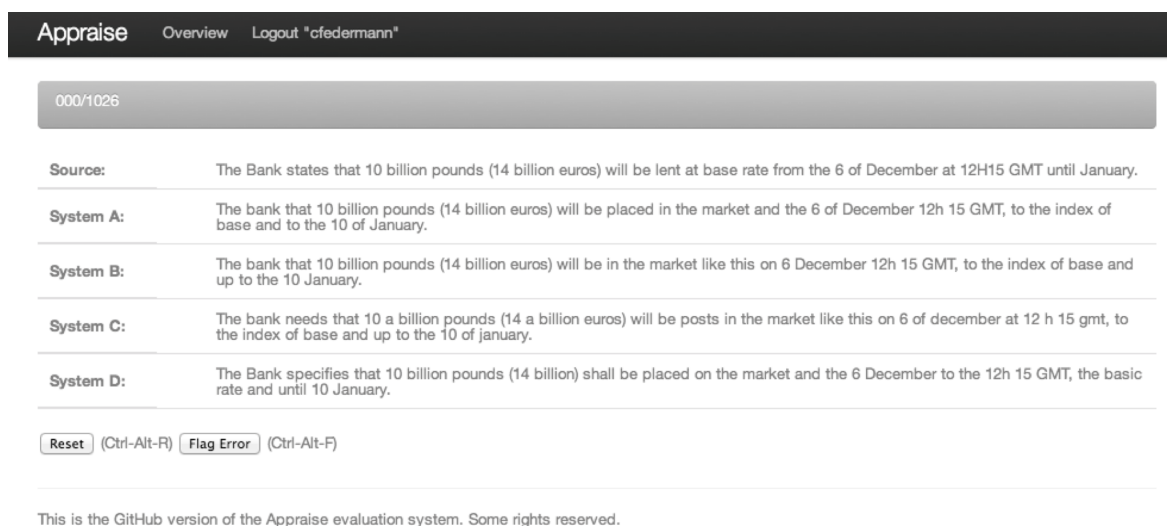


Figure 1: Screenshot of the Appraise evaluation interface.

- Scotland, July. Association for Computational Linguistics.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.
- George Doddington. 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.
- Christian Federmann, Yu Chen, Sabine Hunsicker, and Rui Wang. 2011. DFKI System Combination using Syntactic Information at ML4HMT-2011. 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 (ML4HMT)*, Barcelona, Spain, November. META-NET.
- Christian Federmann. 2010. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta, May. European Language Resources Association.
- Melvyn J. Hunt. 1990. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336.
- Shankar Kumar and William Byrne. 2002. Minimum bayes-risk word alignments of bilingual texts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, EMNLP '02*, pages 140–147, Stroudsburg, PA, USA. Association for Computational Linguistics.
- J.R. Landis and G.G. Koch. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Vladimir Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics Doklady*, volume 10.
- Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.
- Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *In Proceedings of the EACL*, pages 33–40.
- Tsuyoshi Okita and Josef van Genabith. 2011. DCU Confusion Network-based System Combination for ML4HMT. 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 (ML4HMT)*, Barcelona, Spain, November. META-NET.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.
- Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. 2010. Matrex: the dcu mt system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT '10*, pages 143–148, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio

- Ortiz-Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2006. Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, November.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Matthew G. Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23:117–127, September.
- C. Tillmann, S. Vogel, H. Ney, H. Sawaf, and A. Zubiga. 1997. Accelerated dp-based search for statistical translation. In *European Conference on Speech Communication and Technology*, volume 5, pages 2667–2670, Rhodes, Greece, September.
- Vincent Vandeghinste, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, Olga Yannoutsou, Toni Badia, Maite Melero, Gemma Boleda, Michael Carl, and Paul Schmidt. 2008. Evaluation of a machine translation system for low resource languages: METIS-II. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, May.

	Automated Metrics					
	BLEU	NIST	METEOR	PER	WER	TER
DCU	<b>25.32</b>	<b>6.74</b>	56.82	<b>60.43</b>	<b>45.24</b>	<b>0.65</b>
DFKI-A	23.54	6.59	54.30	61.31	46.13	0.67
DFKI-B	23.36	6.31	<b>57.41</b>	65.22	50.09	0.70
LIUM	24.96	6.64	55.77	61.23	46.17	0.65

Table 1: Automated metrics’ scores for ML4HMT test set.

	Automated Metrics					
	BLEU	NIST	METEOR	PER	WER	TER
Joshua	19.68	6.39	50.22	47.31	62.37	<i>n/a</i>
Lucy	<b>23.37</b>	6.38	<b>57.32</b>	49.23	64.78	<i>n/a</i>
Metis	12.62	4.56	40.73	63.05	77.62	<i>n/a</i>
Apertium	22.30	6.21	55.45	50.21	64.91	<i>n/a</i>
MaTrEx	23.15	<b>6.71</b>	54.13	<b>45.19</b>	<b>60.66</b>	<i>n/a</i>

Table 2: Automated metrics’ scores for baseline systems on ML4HMT test set.

System	Annotators			Overall
	#1	#2	#3	
DCU	2.44	2.61	2.51	2.52
DFKI-A	2.50	2.47	2.48	2.48
DFKI-B	<b>2.06</b>	<b>2.13</b>	<b>1.97</b>	<b>2.05</b>
LIUM	2.89	2.79	2.93	2.87

Table 3: Average rank per system per annotator from manual ranking.

System	Ranked				Mode
	1st	2nd	3rd	4th	
DCU	62	79	<b>97</b>	62	3rd
DFKI-A	73	65	<b>82</b>	80	3rd
DFKI-B	<b>127</b>	84	47	42	1st
LIUM	38	72	74	<b>116</b>	4th

Table 4: Statistical mode per system from manual ranking.

	Systems						Annotators		
	DCU, DFKI-A	DCU, LIUM	DFKI-A, LIUM	DCU, DFKI-B	DFKI-A, DFKI-B	DFKI-B, LIUM	#1, #2	#1, #3	#2, #3
$\pi$ -Score	0.296	0.250	0.352	0.352	0.389	0.435	0.331	0.338	0.347

Table 5: Pairwise agreement (using Scott’s  $\pi$ ) for all pairs of systems/annotators.