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

Christian Federmann
Language Technology Lab
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
cfedermann@dfki.de

Abstract

We describe the ML4HMT shared task which aims to foster research on improved system combination approaches for MT. Participants of the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. We describe the ML4HMT corpus and the annotation format we have designed for it and briefly summarize the participating systems. Using automated metrics scores and extensive manual evaluation, we discuss the performance of the various systems. An interesting result from the shared task is the fact that we observed different systems winning according to the automated metrics and according to the manual evaluation. We conclude by summarising the first edition of the challenge and give an outlook to future work.

1 Introduction

The "Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT" is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. Participants of the challenge are requested to build Hybrid/System Combination systems by combining the output of several MT systems of different types and with very heterogeneous types of metadata information, as provided by the organizers.

The main focus of the shared task is trying to answer the following question: *Could Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

Our research in work package 2 of the META-NET project focuses on the design and development of such advanced combination methods, building bridges 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, including information such as part-of-speech, word alignment, or language model scores. The collected data has been released as a multilingual corpus[1]. Furthermore, we have organised a workshop including a challenge exploiting the ML4HMT corpus[2].

The remainder of this paper is structured as follows: in Section 2 we describe the data given to the shared task participants and give a detailed description of the challenge. Section 3 presents the systems taking part in the challenge before we present and discuss evaluation results in Section 4. We conclude by giving a summary of the ML4HMT shared task and an outlook to future work in Section 5.

2 Challenge Description

The participants are given a bilingual development set, aligned at a sentence level. For each sentence, the corresponding *bilingual data set* contains:

- the source sentence,
- the target (reference) sentence, and

[1] Data package available from http://www.dfki.de/~cfedermann/ML4HMT-data-1.0.tgz

[2] See http://www.dfki.de/ml4hmt/

– the corresponding multiple output translations from 5 different systems, based on different MT approaches.

For the ML4HMT data set we decided to use the following systems: Apertium (Ramírez-Sánchez et al., 2006), Joshua (Li et al., 2009), Lucy (Alonso and Thurmair, 2003), MaTrEx (Penkale et al., 2010), and Metis (Vincent Vandeghinste and Schmidt, 2008)). The output has been annotated with system-internal metadata information derived from the translation process of each of the systems.

2.1 Annotated Data Format

We have developed a new dedicated 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 localization. It was standardized by OASIS in 2002 and its current specification is v1.2 released on Feb-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 localized and the corresponding localized (translated) data for one locale only. The localizable texts are stored in `<trans-unit>` elements each having 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 (in the `"metanet"` namespace) allowing a wide variety of meta-data annotation 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 several parameters (model weights) which are described in the `<metanet:weight>`.

Annotation is stored in `<alt-trans>` element within the `<trans-unit>` elements. The `<source>` and `<target>` elements in the `<trans-unit>` elements refer to the source sentence and its reference translation, respectively. The `<source>` and `<target>` elements in the `<alt-trans>` elements specifies the input and 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>` element. Its content is tool-specific.

The full format specification is available as an XML schema. An example annotation from the ML4HMT data set is depicted in Figure 1.

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 organizers 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 system described in Okita and van Genabith (2011) presents a system combination module in the MT system MaTrEx (Machine Translation using Examples) developed at Dublin City University. A system combination module deployed by them 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. Their system is based on system combination techniques which use a confusion network on top of a Minimum Bayes Risk (MBR) decoder (Kumar and Byrne, 2002).

One interesting, novel point in their submission is that for the given single best translation outputs, they tried to identify which inputs they will consider for the system combination, possibly discarding the worst performing system(s) from the combination input. As a result of this selection process, their BLEU score, from the combination of the four single best systems, achieved 0.48 BLEU points absolute higher than the combination of the five single best systems.

3.2 DFKI-A

A system combination approach with a sentence ranking component is presented in Avramidis (2011). The paper reports on a pilot study on a Hybrid Machine Translation that takes advantage of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting in a selection mechanism able to learn and rank and select systems' translation output on the complete sentence level, based on their respective quality.

For training, due to the lack of human annotations, word-level Levenshtein distance 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 on up to 54% of the cases.

3.3 DFKI-B

The authors of Federmann et al. (2011) report on experiments that are focused on word substitution using syntactic knowledge. From the data provided by the workshop organisers, they choose one system to provide the "translation backbone". The Lucy MT system was suited best for this task, as it offers parse trees of both the source and target side, which allows the authors to identify interesting phrases, such as noun phrases, in the source and replace them in the target language output. The remaining four systems are mined for alternate translations on the word level 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:

- the part-of-speech of the original translation must match the candidate fragment.
- Additionally they may consider the 1-left and 1-right context.
- Besides the part-of-speech, all translations plus their context are scored with a language model trained on EuroParl (Koehn, 2005).
- Additionally, the different systems may turn up with the same translation, in that case the authors select the candidate with the highest count ("majority voting").

The authors reported improvements in terms of BLEU score when comparing to the translations from the Lucy RBMT system.

3.4 LIUM

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

1. 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.

2. The second module is the decoder. This decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The costs computed in the decoder can be expressed as a weighted sum of the logarithm of feature functions. The following features are considered in decoding:
   - the language model probability, given by a 4-gram language model
   - a word penalty, which depends on the number of words in the hypothesis
   - a null-arc penalty, which depends on the number of null arcs crossed in the lattice to obtain the hypothesis
   - the system weights: each word receives a weight corresponding to the sum of the weights of all systems which proposed it.

4 Evaluation Results

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

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

Table 1: Automated scores for ML4HMT test set.

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 2. 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 Carletta (1996) and compute κ scores to estimate the inter-annotator agreement. In our manual evaluation campaign, we had $n = 3$ annotators so computing basic, pairwise annotator agreement is not sufficient—instead, we apply Fleiss (1971) who extends Scott (1955) for computing inter-annotator agreement for $n > 2$.

Annotation Setup As we have mentioned before, 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.*)[3]. In a second evaluation scenario, we only collected the *1-best* ranking system per sentence, resulting in a total of four categories (A: "system A ranked 1st", etc.). In this second scenario, we can expect a higher annotator agreement due to the reduced number categories. Overall, we collected 904 sentences with an overlap of $N = 146$ sentences for which all annotators assigned ranks.

Scott's π allows to measure the pairwise annotator agreement for a classification task. It is defined as

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

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 π score of

$$\pi = \frac{0.673 - 0.5}{1 - 0.5} = 0.346 \tag{2}$$

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". We have decided to compute Scott's π scores to be comparable to WMT11 (Bojar et al., 2011).

Fleiss κ Next to the π scores, there also exists the so-called κ score. Its basic equation is strikingly

[3]Given this huge number of possible categories, we were already expecting resulting κ scores to be low.

System	BLEU	NIST	METEOR	PER	WER
Joshua	19.68	6.39	50.22	47.31	62.37
Lucy	**23.37**	6.38	**57.32**	49.23	64.78
Metis	12.62	4.56	40.73	63.05	77.62
Apertium	22.30	6.21	55.45	50.21	64.91
MaTrEx	23.15	**6.71**	54.13	**45.19**	**60.66**

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

System	Annotator #1	Annotator #2	Annotator #3	Overall
DCU	2.44	2.61	2.51	2.52
DFKI-A	2.50	2.47	2.48	2.48
DFKI-B	**2.06**	**2.13**	**1.97**	**2.05**
LIUM	2.89	2.79	2.93	2.87

Table 3: Average rank per system per annotator from manual ranking of 904 (overlap=146) translations.

similar to (1)

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (3)$$

with the main difference being the κ score's support for $n > 2$ annotators. We compute κ for two configurations. Both are based on $n = 3$ annotators and $N = 146$ sentences. They differ in the number of categories that a sentence can be assigned to (k)

1. *complete* scenario: $k = 24$ categories. For this, we obtained an overall κ score of

$$\kappa_{complete} = \frac{0.1 - 0.054}{1 - 0.054} = 0.049 \quad (4)$$

2. *1-best* scenario: $k = 4$ categories. For the reduced number of categories, κ improved to

$$\kappa_{1-best} = \frac{0.368 - 0.302}{1 - 0.302} = 0.093 \quad (5)$$

It seems that the large number of categories of the *complete* scenario has indeed had an effect on the resulting $\kappa_{complete}$ score. This is a rather expected outcome, still we report the κ scores for future reference. The *1-best* scenario supports an improved κ_{1-best} score but does not reach the level of agreement observed for the π score.

It seems that DFKI-B was underestimated by BLEU scores, potentially due to its rule-based characteristics. This is a possible reason for the relatively higher inter-annotator agreement when compared with other systems. Also, DCU and LIUM may have low inter-annotator agreement as their background is similar.

Due to the fact that κ is not really defined for *ordinal data* (such as rankings in our case), we will investigate other measures for inter-annotator agreement. It might be a worthwhile idea to compute α scores, as described in Krippendorff (2004). Given the average rank information, statistical mode, π and κ scores, we still think that we have derived enough information from our manual evaluation to support for future discussion.

5 Conclusion

We have developed an Annotated Hybrid Sample MT Corpus which is a set of 2,051 sentences translated by five different MT systems[4] (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 metric scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly,

[4]Not all systems available for all language pairs.

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

Table 4: Statistical mode per system from manual ranking of 904 (overlap=146) translations.

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

Table 5: Pairwise agreement (using Scott's π) for all pairs of systems/annotators.

the system winning nearly all the automatic scores (DCU) only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings (DFKI-B) 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 wrt. the evaluation of such systems' output, needs to be undertaken. We have learned from the participants that our ML4HMT corpus is 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. We are looking forward to an interesting workshop and thank the participants for their efforts during the ML4HMT-2011 Shared Task.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119). We thank the organisers of LIHMT 2011 for their support.

References

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium 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. to appear.

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.

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, 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.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22:249–254, June.

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. to appear.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *LREC*.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Klaus Krippendorff. 2004. Reliability in content analysis. some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

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.

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.

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. to appear.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

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.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, 19(3):321–325.

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.

Ineke Schuurman Stella Markantonatou Sokratis Sofianopoulos Marina Vassiliou Olga Yannoutsou Toni Badia Maite Melero Gemma Boleda Michael Carl Vincent Vandeghinste, Peter Dirix and Paul Schmidt. 2008. Evaluation of a machine translation system for low resource languages: Metis-ii. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

```
<trans-unit id="s71">
  <source xml:lang="es">El paciente fue aislado.</source>
    <target xml:lang="en">The patient was isolated.</target>
    <alt-trans rank="1" tool-id="t3">
      <source xml:lang="es">El paciente fue aislado.</source>
      <target xml:lang="en">The paciente was isolated .</target>
      <metanet:scores>
        <metanet:score type="total" value="-60.4375047559049"/>
      </metanet:scores>
      <metanet:derivation id="s71_t3_r1_d1">
        <metanet:phrase id="s71_t3_r1_d1_p1">
          <metanet:string>The</metanet:string>
          <metanet:annotation type="lemma" value="the"/>
          <metanet:annotation type="pos" value="AT0"/>
          <metanet:annotation type="morph_feat" value=":m:sg:"/>
          <metanet:alignment from="0" to="0"/>
        </metanet:phrase>
```

Figure 1: Example of annotation from the ML4HMT corpus.

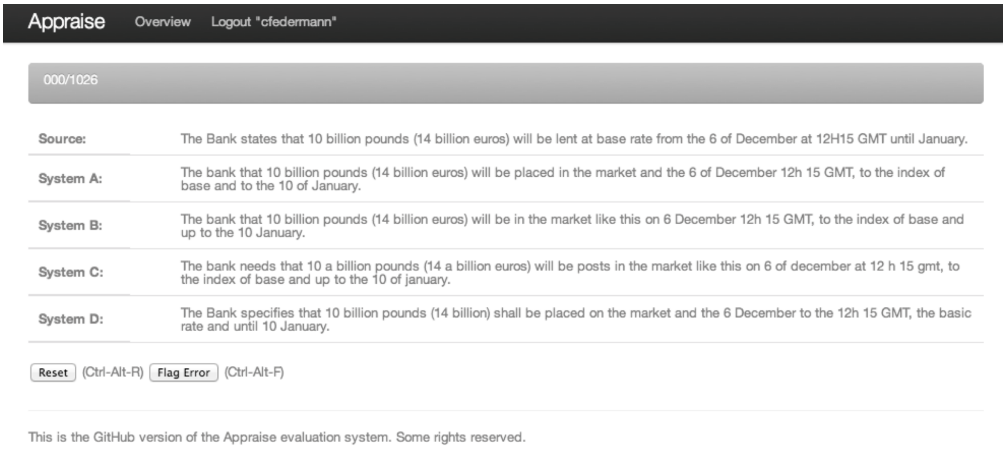

Figure 2: Screenshot of the Appraise interface for human evaluation.