Results from the ML4HMT Shared Task

Christian Federmann, DFKI LT Lab

Overview

- Introduction
- Challenge Description
- Participating Systems
- Evaluation Results
- Conclusion

Introduction

Introduction

- ML4HMT shared task is "*an effort to trigger systematic investigation on improving hybrid MT*"
- Special focus on machine learning techniques
- Participants are requested to build hybrid translations with the ML4HMT corpus data
- *"Could hybrid MT techniques benefit from extra information from the different systems?"*

Motivation

- WP2 from META-NET focuses on building bridges to the machine learning community
- Joint and systematic exploration of novel system combination approaches
- For this, we released a multilingual corpus
- And organised the ML4HMT workshop

Challenge Description

ML4HMT Data

- Bilingual development set, sentence-aligned
- Available information
  - source, target (reference)
  - translation output from 5 MT systems
- Translation output may contain annotations
- Systems: Apertium, Joshua, Lucy, MaTrEx, Metis

ML4HMT Data, cont'd

- Translation output annotated with metadata
- Annotated data format derived from XLIFF
- WMT 2008 test set used as source text
- development set 1,025 sentences
- test set 1,026 sentences

ML4HMT Data, cont'd

```xml
<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>
```

Participating Systems

Participating Systems

- DCU, Okita and van Genabith
- DFKI-A, Avramidis
- DFKI-B, Federmann et al.
- LIUM, Barrault and Lambert

Evaluation Results

Evaluation Setup

- Automated scores
  - BLEU, NIST, METEOR, PER, WER, TER
- Extensive manual ranking evaluation
  - 3 annotators ranking 904 sentences
  - Overlap of 146 sentences

Baseline Scores

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**

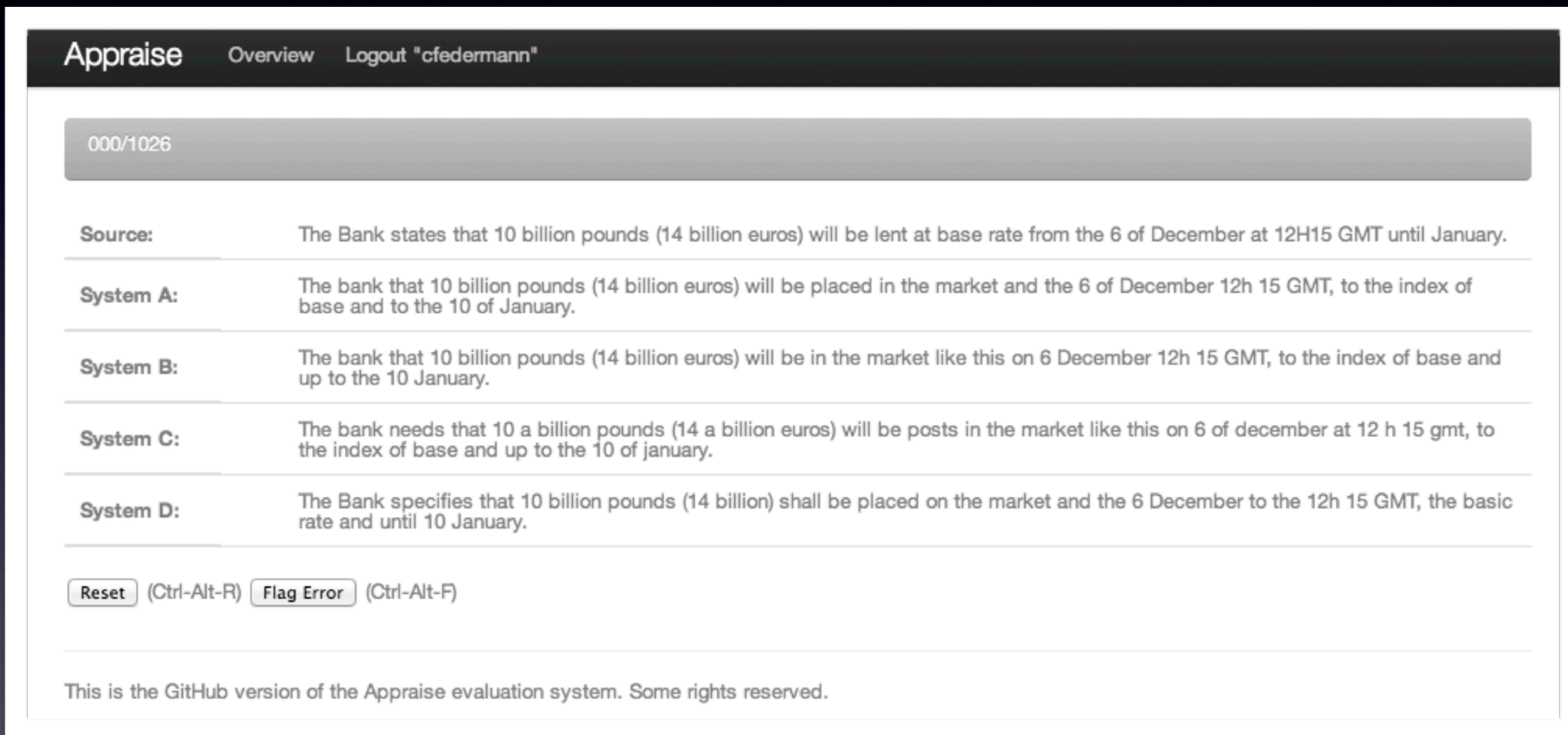
Automated Scores

System	BLEU	NIST	METEOR	PER	WER	TER
DCU	**25.52**	**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

Manual Evaluation

- We used the Appraise evaluation system
- Users see a reference and four translations
- These are then ranked in *best-to-worst* order

Evaluation Interface



Manual Ranking

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

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

Manual Ranking, cont'd

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

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

Annotator Agreement

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

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

Findings

- "*fair agreement*" according to Scott's π
- Fleiss' κ scores affected by many categories
  - we need more annotators next time
  - also, different metrics could be applied?
  - even simpler1-best scenario did not help
- Overall results: DCU wins by automated scores, DFKI-B in the manual evaluation

Conclusion & Outlook

Conclusion

- Created an annotated corpus for hybrid MT
- Using this resource, we have setup ML4HMT
- 4 participating systems, different approaches
- Interesting results: automated vs. manual
- All participating systems improved over the baseline systems!

Outlook

- We have to improve the ML4HMT data set
- The shared task description also needs some re-writing to avoid confusing participants
- Evaluation of combo results is challenging
- Further investigation of hybrid combination methods is needed
- Attract more machine learning researchers!

Thanks for joining!

- We would like to thank all participants from the shared task and the workshop
- Thanks to Maite, Marta, and Toni for taking on the organisational details here in Barcelona
- We hope to see you again at ML4HMT-12
- Any feedback you have is greatly appreciated!

Thank you!

Questions & Answers

Acknowledgements

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 for their support.