ML4HMT-2012

The Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation

COLING 2012

Workshop and Shared Task

Organizers:

- Christian Federmann (German Research Center for Artificial Intelligence (DFKI))
- Dr. Maite Melero (Barcelona Media (BM))
- Dr. Marta R. Costa-jussà (Barcelona Media (BM))
- Prof. Toni Badia (Universitat Pompeu Fabra and Barcelona Media (BM))
- Dr. Tsuyoshi Okita (Dublin City University (DCU))
- Prof. Josef van Genabith (Dublin City University (DCU) and Centre for Next Generation Localisation (CNGL))

Machine Translation Paradigms

- RB-MT – Rule-Based Machine translation
- EB-MT – Example-Based Machine Translation
- SMT – Statistical Machine Translation
- PB-SMT – Phrase-Based Statistical Machine Translation
- HPB-SMT – Hierachical Phrase-Based Statistical Machine Translation
- SB-SMT – Syntax-Based Statistical Machine Translation

...

Observation: Different systems have different strengths

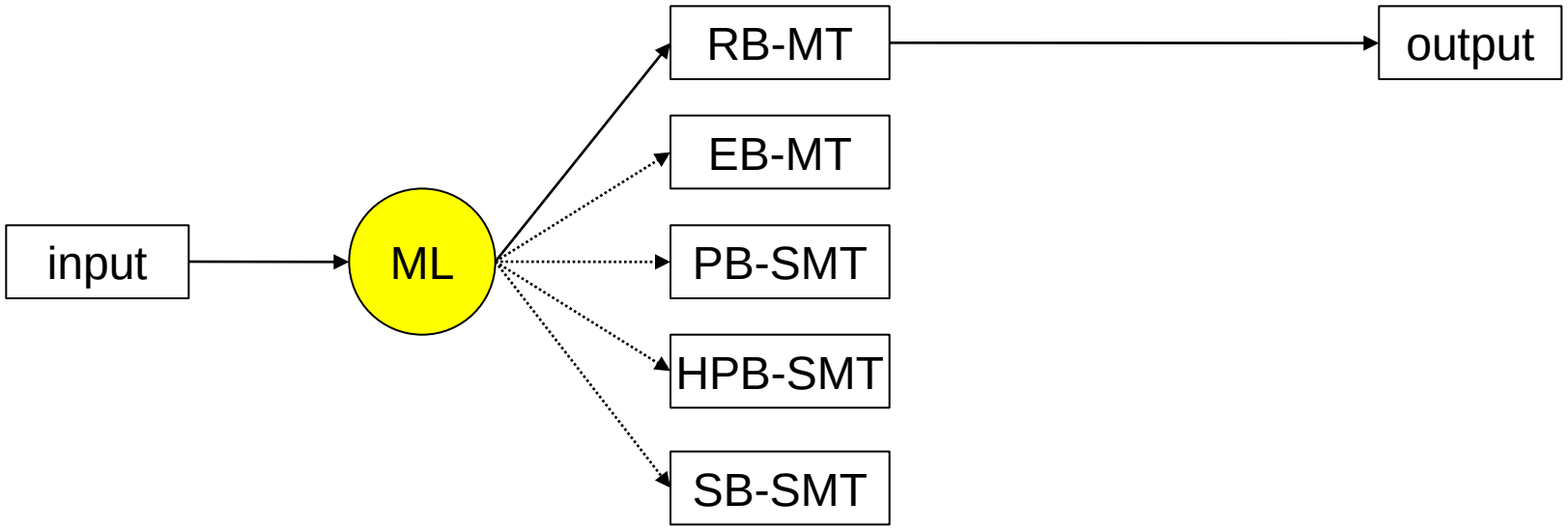
(e.g. easy training of SMT vs. good grammar of RB-MT)

Hypothesis: can hybrid systems combine best of all?

How: Machine Learning

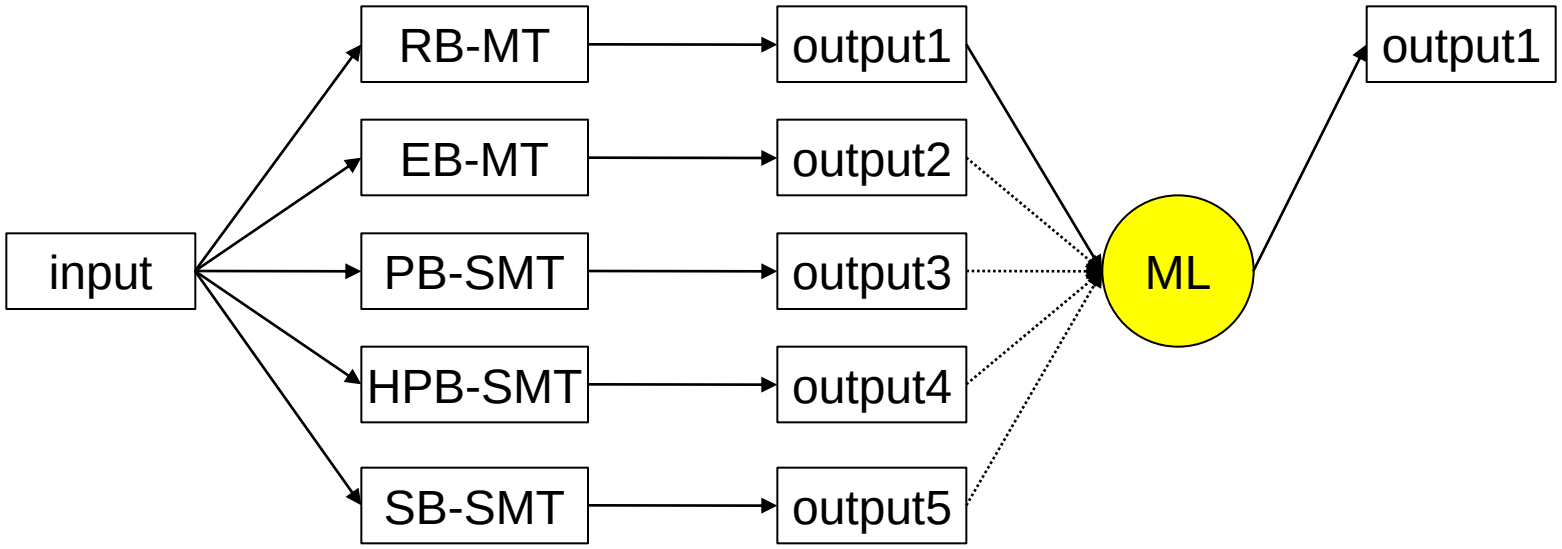
Hybrid MT: Pre-Translation System Selection

- multiple MT engines/systems available
- machine learning techniques to decide which system is best to translate input sentence



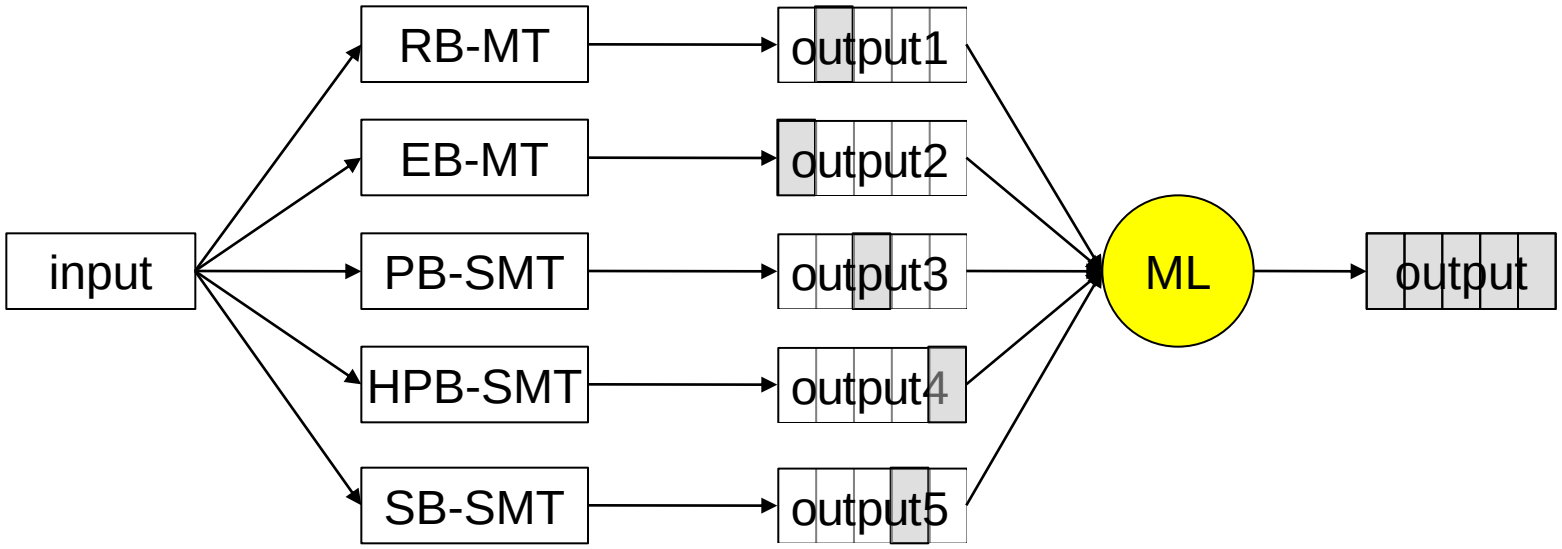
Hybrid MT: Post-Translation System Selection

- multiple MT engines/systems available
- the input sentence is translated by all the systems and the best translation is selected based on the analysis of their ouptuts



Hybrid MT: Post-Translation System Selection

- multiple MT engines/systems available
- all systems are used to produce multiple translations of the input sentence, they are broken down to smaller pieces and these are recombined to get a better output



ML4HMT-2012

- Based on system combination:
- Multiple systems based on different paradigms used to produce annotated n-best outputs:
  - **MaTrEx (example based):** all language pairs ↔ English
  - **Moses (phrase based):** all language pairs ↔ English
  - **Metis (rule based):** Spanish → English, German → English
  - **Apertium (rule based):** Spanish ↔ English
  - **Lucy (rule based):** Spanish, German ↔ English
  - **Joshua (hierarchical phrase based):** all language pairs ↔ English
  - **TectoMT (deep syntax based):** Czech ↔ English
- **Annotation:** words, phrases, subtrees, chunks scored by different models (depending on the system)
- **Decoding:** machine learning to use strings + meta-data for better output

- Many more
  - Statistical post-editing: RBMT > SMT or SMT > SMT
  - Pre-ordering
  - ….
- System combination:
  - parallel, sequential, …, but not just MEMT
  - probabilities in RBMT etc.

Optimising the Division of Labour in Hybrid MT

- **Objectives:** To provide a systematic investigation and exploration of the space of possible choices in Hybrid MT, in order to provide optimal support for Hybrid MT design, using sophisticated machine-learning (ML) technologies.
- **Partners:**
  - **DFKI** – Deutsche Forschungszentrum für Künstliche Intelligenz (Germany)
  - **BM** – Barcelona Media (Spain)
  - **DCU** – Dublin City University (Ireland)

ML4HMT 2012

- Barcelona, Spain 2011
- Mumbai, India 2012 COLING
  - A very hard task
  - Heterogeneous and often incompatible meta-data
  - Difficult for ML
- WS a combination
  - Regular papers
  - Shared task
- Wider focus on general machine learning for/in MT

9:00 Josef van Genabith - Welcome and introductory remarks

Regular Papers:

9:15 **Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation**
Vassilina Nikoulina, Agnes Sandor, Marc Dymetman

9:40 **Confusion Network Based System Combination for Chinese Translation Output: Word-Level or Character-Level?** Maoxi Li, Mingwen Wang

10:05 **Using Cross-Lingual Explicit Semantic Analysis for Improving Ontology Translation**
Kartik Asooja, Jorge Gracia, Nitish Aggarwal, Asunción Goméz Pérez, presented by Mihael Arcan

Shared Task

10:30 **System Combination with Extra Alignment Information**
Xiaofeng Wu
10:50 **Topic Modeling-based Domain Adaptation for System Combination**
Antonio Toral
11:10 **Sentence-Level Quality Estimation for MT System Combination**
Raphaël Rubino
11:30 Tea break
11:45 **Neural Probabilistic Language Model for System Combination**
Tsuyoshi Okita
12:05 **System Combination Using Joint, Binarised Feature Vectors**
Christian Federmann

12:25 **Results of the ML4HMT-12 Shared Task**
Christian Federmann, Tsuyoshi Okita, Maite Melero, Marta Ruiz Costa-Jussà, Toni Badia, Josef van Genabith

12:30 **Discussion Panel**

Panelists: Jan Hajič, Qun Liu, Hans Uszkoreit, Josef van Genabith

Topics include:

- The Future of Hybrid MT: is there a single-paradigm winner?
- Will we see increasing usage of additional, potentially highly sparse, features?
- Will research efforts in Machine Translation and Machine Learning converge?
- How do we evaluate progress in terms of translation quality for Hybrid MT?
- What are the baselines? Can Human Judgment be integrated?

12:50 Invited talk: **Deep Linguistic Information in Hybrid Machine Translation**

Jan Hajič · Institute of Formal and Applied Linguistics · Charles University in Prague

13:30 Lunch