Machine Translation Research in META-NET

Jan Hajič

Institute of Formal and Applied Linguistics
Charles University in Prague, CZ
hajic@ufal.mff.cuni.cz
With contributions by Marcello Federico, Pavel Pecina, Stephan Peitz and Timo Honkela

META-FORUM 2010: Challenges for Multilingual Europe
Brussels, Belgium, November 17/18, 2010

Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

Outline

- **Pillar I in META-NET**
  - …the research element of META-NET
- **Semantics in Machine Translation**
  - Semantic features in statistical MT
  - (Semantic) Tree-based translation
- **Hybrid MT systems**
  - Rule-based and statistical
- **Context in MT**
  - „Extra-linguistic“ features
- **More data for MT**
  - Parallel data for under-resources langauges
- **Related projects & the Future**

Semantics in Machine Translation

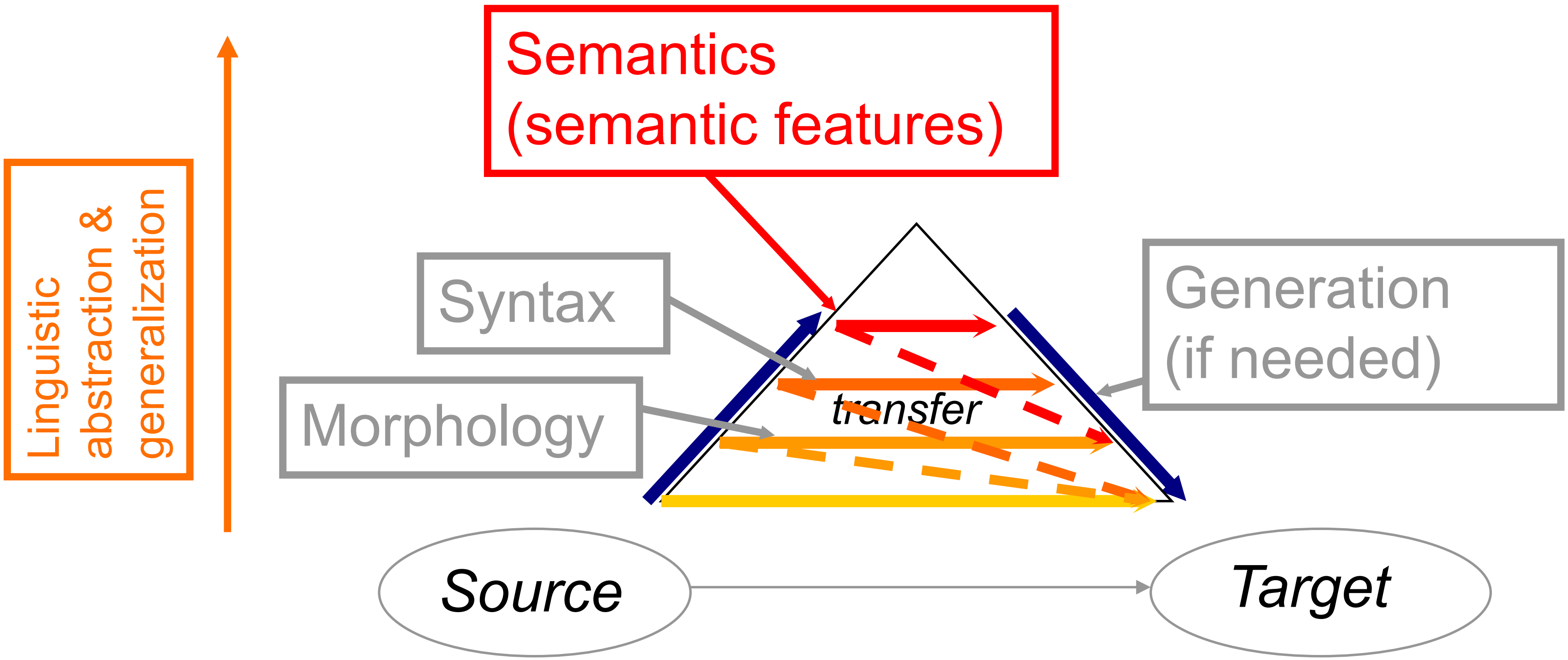
Semantics in Machine Translation

META-NET

- **What is semantics, anyway?**
  - For now: anything beyond and outside morphology and syntax
    - Semantic Roles (words vs. predicates)
    - Lexical Semantics (WSD), MWE
    - Named Entities
    - Co-reference (pronominal, bridging anaphora)
    - Textual Entailment
    - Discourse Structure
    - Information Structure … + any combination of the above
- **New metrics**
  - BLEU, METEOR, NIST etc. biased towards (good) local n-grams
  - Metrics sensitive to semantics?
- **Tools and Resources**
  - Semantically annotated parallel corpora; metrics tools, analysis tools

Semantics in Machine Translation

- **Analysis – transfer [– generation]**



Semantics in Machine Translation

- **Case Study 1**
  - Cross-lingual Textual Entailment for Adequacy Evaluation

    Y. Mehad, M. Negri, M. Federico: *Towards cross-lingual textual entailment*, NAACL 2010
- **Case Study 2**
  - Combined Syntax and Semantics for MT Transfer

    D. Mareček, M. Popel, Z. Žabokrtský: *Maximum Entropy Translation Model in Dependency-Based MT Framework*, WMT / ACL 2010
- **Case Study 3**
  - Anaphora Resolution for translation of pronouns

    C. Hardmeier, M. Federico: *Modeling Pronominal Anaphora in Statistical MT*, IWSLT 2010.
- **Case Studies → Selected Challenges**
  - Evaluation of impact of individual additions
    - Evaluation data with/without phenomenon under study
    - Automatic vs. human evaluation

Hybrid MT Systems

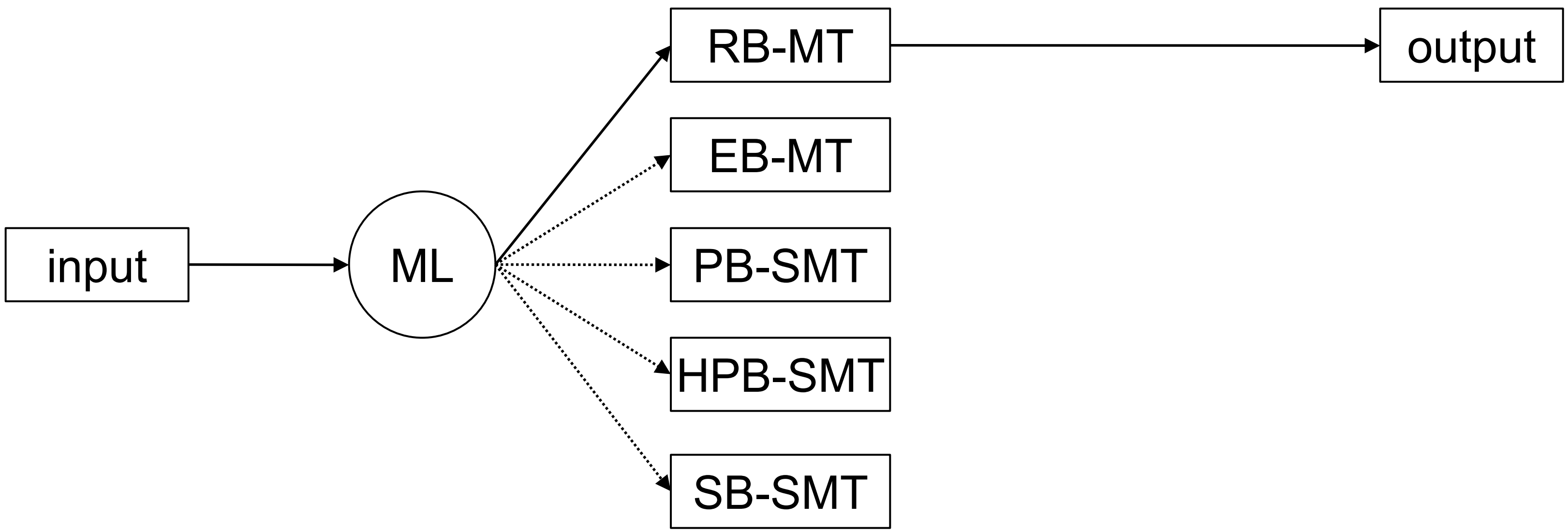
Machine Translation Paradigms

- 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:** Hybrid systems can combine best of all

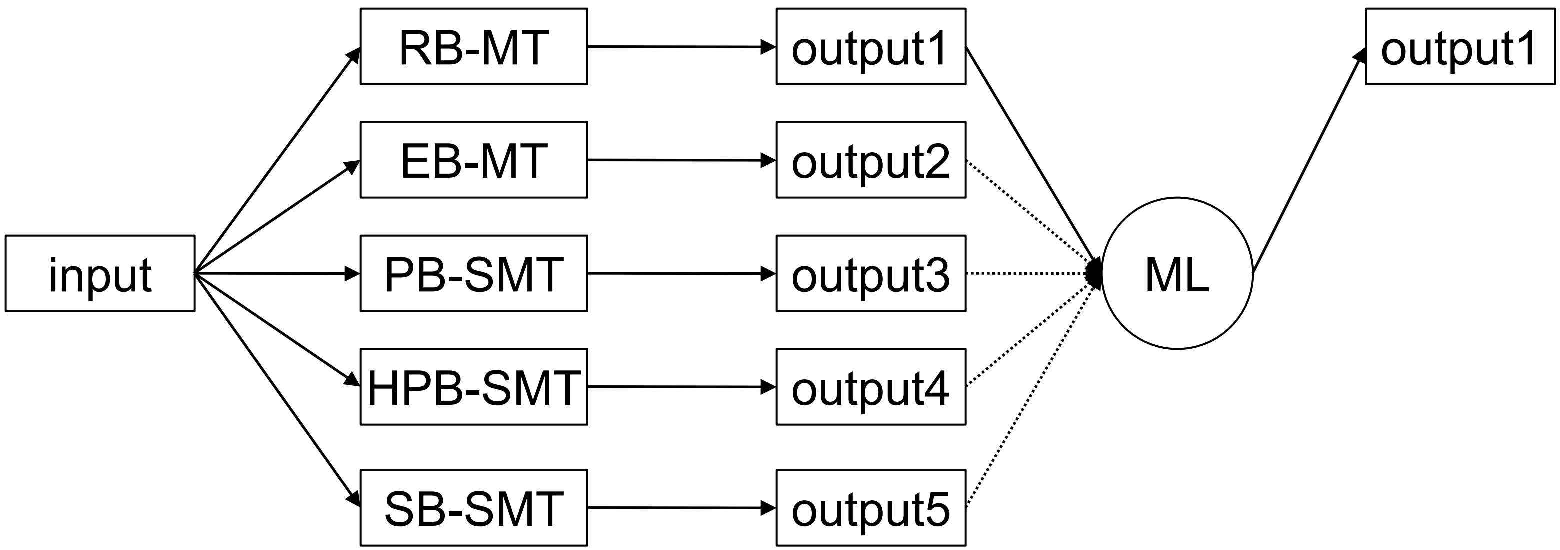
Hybrid MT: Pre-Translation System Selection

- **Multiple MT engines/systems available**
- **Machine learning techniques**
  - decide which system is best to translate the input sentence



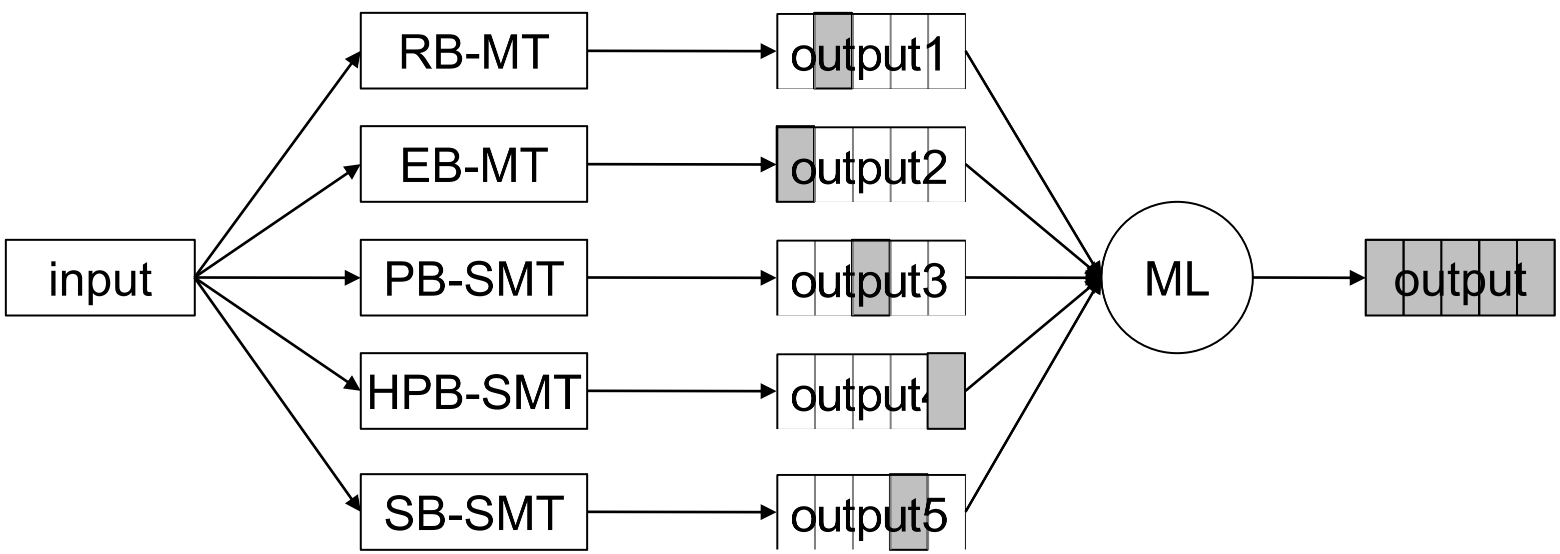
Hybrid MT: Pre-Translation System Selection

- **Multiple MT engines/systems available**
- **All systems translate**
  - Analysis of ouptuts → select translation



Hybrid MT: Pre-Translation System Selection

- **Multiple MT engines/systems available**
- **All systems translate**
  - Translation compiled from analyzed pieces



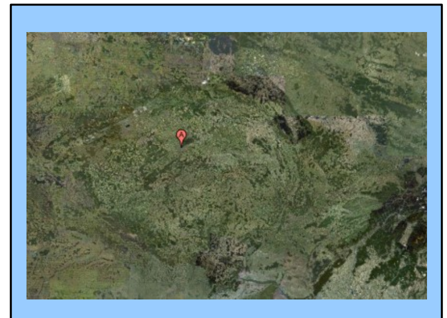
The META-NET Hybrid System Approach

- 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 techniques used to recombine those to get better output

Context in Machine Translation

Increase MT quality and services in multimodal context

(CONTEXTS)



(SOURCE)

Česká republika je jedním z mála vnitrozemských států, jehož obrysy lze rozeznat na satelitních snímcích.

MT

(TARGET)

Czech Republic is one of the few inland countries whose borders can be seen from satellite photographs.

Context in Machine Translation

- **Domain adapted language and translation models**
  - Method
    - Large corpus divided in predefined domains
    - Train translation and language models on each domain
    - Train additional language models on the predefined domains
    - Train a classifier to classify incoming documents to a domain
    - Decode using respective translation and language models
    - Evaluate results and revise method if necessary
  - Resources
    - JRC-Acquis & Eurovoc
    - Europarl
  - Innovation
    - Design, implement and fine-tune classification algorithms
    - Explore ways to effectively combine language and translation models

Context in Machine Translation

- **Context in statistical morphology learning**
  - *O. Kohonen, S. Virpioja, L. Leppänen and K. Lagus (2010): Semisupervised Extensions to Morfessor Baseline*
- **Multimodal context in translation**
  - Research questions:
    - Which kind of multimodal contextual information can be used to advance MT quality? How to better access multimodal information?
    - In which MT applications multimodal information is useful?
  - Current target: enhancing language and translation models with visual and textual context data and ontological knowledge
    - Use cases: translation of figure captions, translation of subtitles, MT in extended reality applications, robotics applications

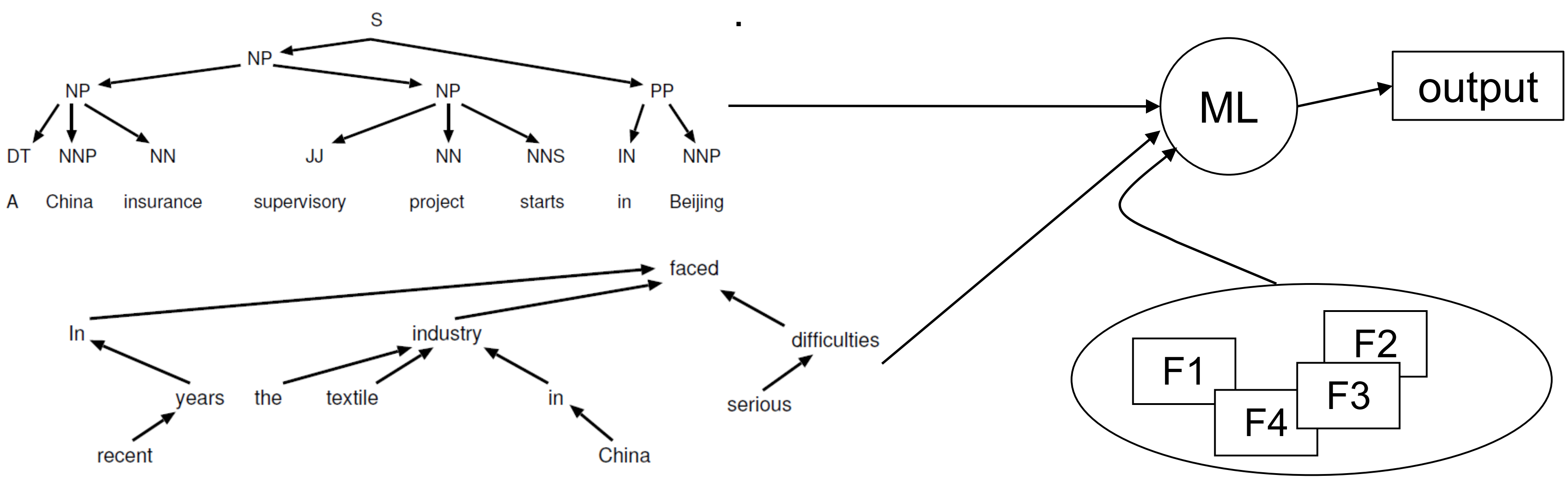
Context in Machine Translation: 2011 Challenge

- **Data**
  - JRC Acquis corpus, 22 European languages
  - Translations by the state-of-the-art statistical systems
- **Tasks**
  - To choose to the best translation from a set candidate translations by multiple systems (reranking task)
  - Context is given by the source sentence, larger linguistic context and the domain of the text
- **Goals**
  - To discover the set of best context features, find representation
  - To foster collaboration between MT and Machine Learning (ML) researchers; infuse MT research with advances from the ML field
- **Future Challenge: 2013**
  - Using visual context (images)

Data and Machine Learning for MT

Data and Advanced Machine Learning in MT

- **“There is no data like more data”**
  - Data crawling, cleanup, deduplication, …
  - Available through META-SHARE
- **Advanced Machine Learning Experiments**
  - Combining several previously described approaches



Related Projects

EU 7th FP Machine Translation (selected projects)

- **EuromatrixPlus**
  - Machine Translation in general – now 8 selected languages (Czech, English, French, Spanish, German, Italian, Slovak, Bulgarian)
- **FAUST**
  - Improving fluency, incorporating user feedback (fast)
  - French, English, Czech, Spanish
- **ACCURAT**
  - Using comparable corpora, esp. for low-resource languages
  - Estonian, Croatian, …
- **LetsMT! (PSP)**
  - Building of data resources (low-resourced languages)
  - For business and research
- **Panacea**
  - Building Resources & Language Tools
  - Tools + Resources → Automatically analyzed corpora
- **Khresmoi (IP)**
  - Medical information retrieval for patients and practitioners
  - Cross-language (English, German, Czech, French) ← MT

The Future

The Future

- **Resources, resources, resources**
  - … and their avialabilty (META-SHARE)
- **Novel, high-risk research**
  - Linguistics
    - Unclear "which linguistics", but some
  - Language Understanding
    - Context, domain knowledge (ontologies?), other modalitites
  - …but SMT is here to stay (in some form)
    - … even though we might not recognize the current "kitchen-sink" paradigm a few years from now
  - New algorithms
    - Neural networks (finally?), Genetic algorithms, Brain research, …
  - Better [automatic] evaluation to guide progress
- **Commercial Applications**
  - Post-editing (CAT) tools with integrated (S)MT, novel features, ergonomics
  - Multilingual information access, information extraction, summarization, sentiment

Q/A

META NET

Thank you very much.

office@meta-net.eu

http://www.meta-net.eu

http://www.facebook.com/META.Alliance