Machine Translation Research in META-NET

Jan Hajič

Institute of Formal and Applied Linguistics
Charles University in Prague, CZ
hajic@ufal.mff.cuni.cz

META FORUM 2011 Solutions for Multilingual Europe
Budapest, Hungary, June 27/28 eu 2011.hu

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

Outline

- **Pillar I in META-NET**
  - …the research element of META-NET
- **Results in the first year**
  - Semantics in Machine Translation
  - Hybrid MT systems
  - Context in MT
  - More data for MT
- **Related projects & the Future**
- **META-NET Challenges in 2012**

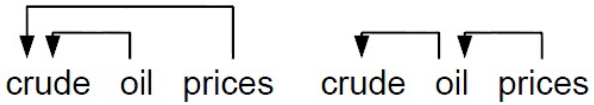
Semantics in Machine Translation

Semantics in Machine Translation

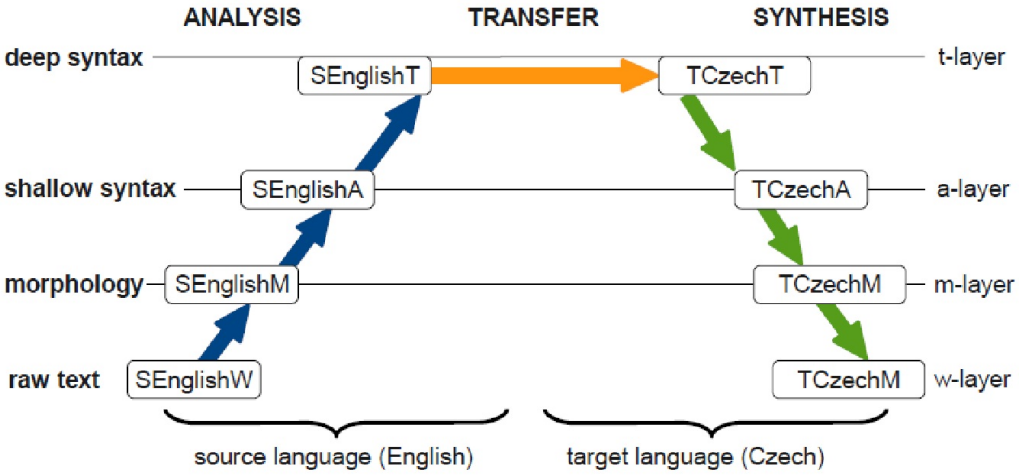
- Anaphora resolution
  - [The same **hospital**]1 had had to contend with a similar infection early this year.[**It**]2→1 had discharged a patient admitted after a serious traffic accident. Shortly afterward, [**it**]3→2 had to re-admit the patient because of an MRSA infection, and [**doctors**]4 have been unable to perform surgery that would be vital to full recovery because [**they**]5→4 have been unable to get rid of the staph.
  - BART-based anaphora resolution + gender prediction (Eng->German):
    - p(**'es'** |neut,sg)= 0.9 p(**'er'** |neut,sg)= 0.05 p(**'sie'** |neut, sg)= 0.02
  - Full system at WMT 2011

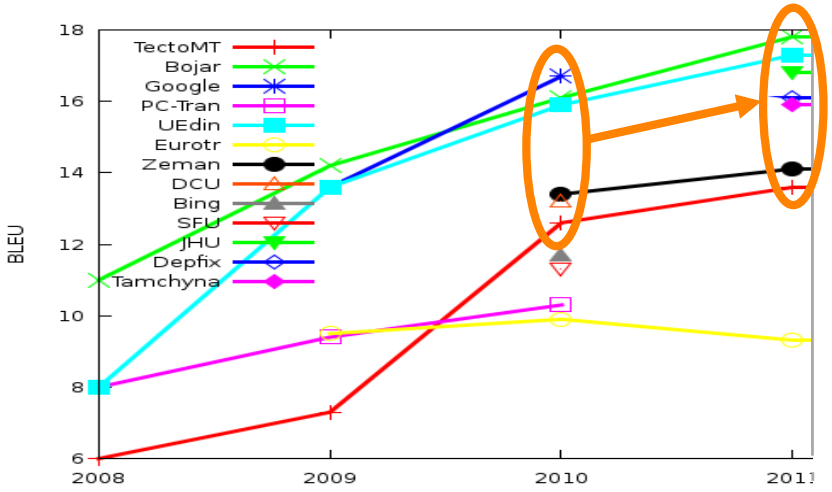
Semantics in Machine Translation

- "Back" to the traditional
  - A – T- G model
- Most steps
  - Statistical model
- NP structure parsing



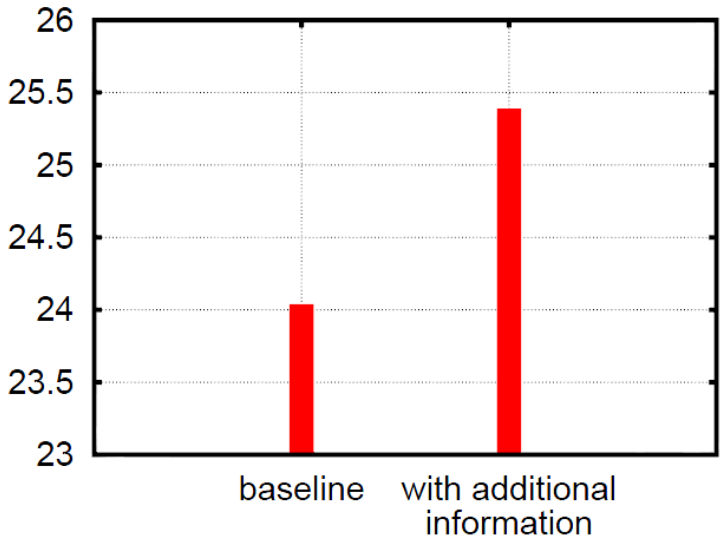
  - MT: ~0.6 BLEU points
- Generation
  - Improves translation to *inflective languages*
- Improving in fastest pace
  - WMT'2011 (Edinburgh)

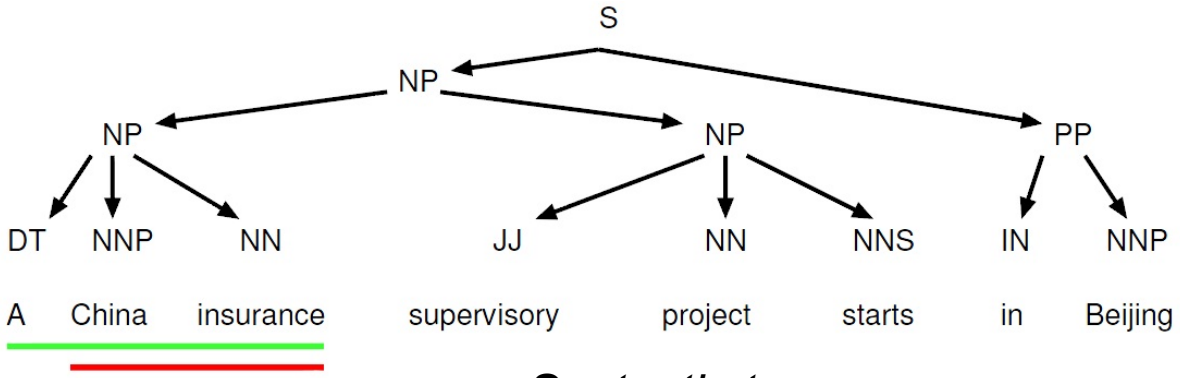




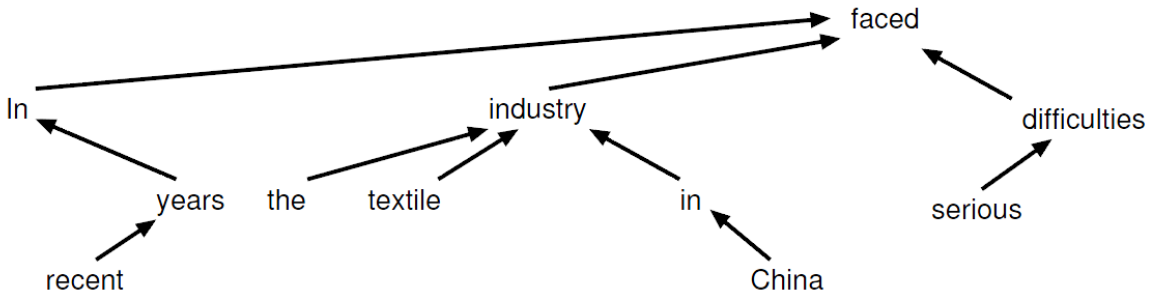
Semantics in Machine Translation

- Syntax/Semantics on target side only
- Significant BLEU improvement
- English target side only





Syntactic tree

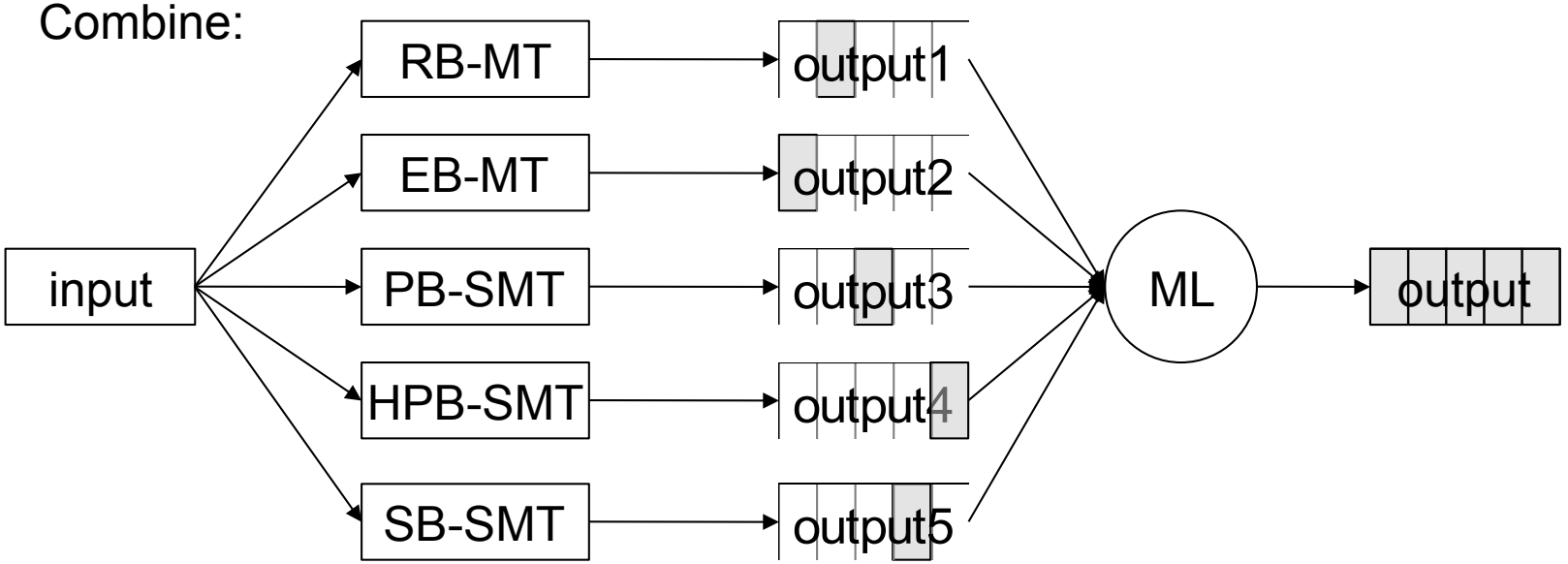


Dependency ("semantic") tree

Hybrid MT Systems

Hybrid MT Systems

- Multiple systems, different technologies → different results
  - Combine:



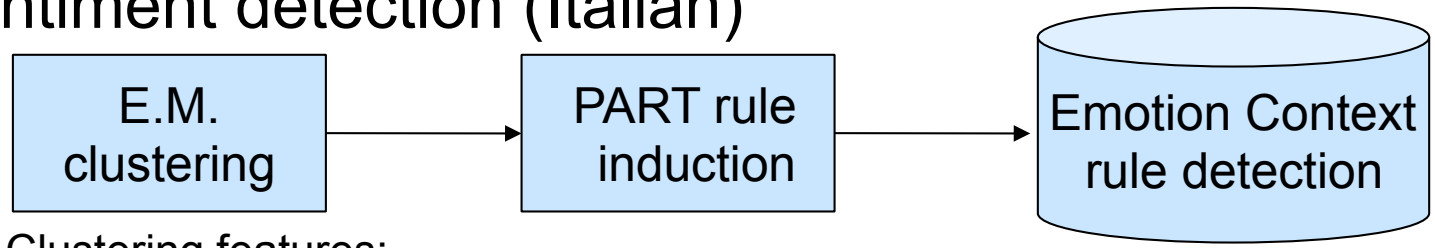
- Bottleneck: data availability…
- …data now available:
  - 5 systems, 3 language pairs (Eng/Spa, Eng/Cze), rich annotation, scores

DCU, BM, DFKI

Context in Machine Translation

Context in Machine Translation

- Topic adaptation
  - Topic detection: POLITICS, GEOGRAPHY, LAW, FINANCE, …
  - Unsupervised (bilingual Probabilstic Latent Semnatic Analysis)
    - Spoken translation (TED, Eng-Fre)
    - 2.4-3% improvement BLEU/NIST score, lower LM perplexity
  - Supervised
    - Eurovoc top-level domains, EuroParl corpus
    - BLEU increase by 0.15-0.2 for all domains with sufficient data
- Sentiment detection (Italian)

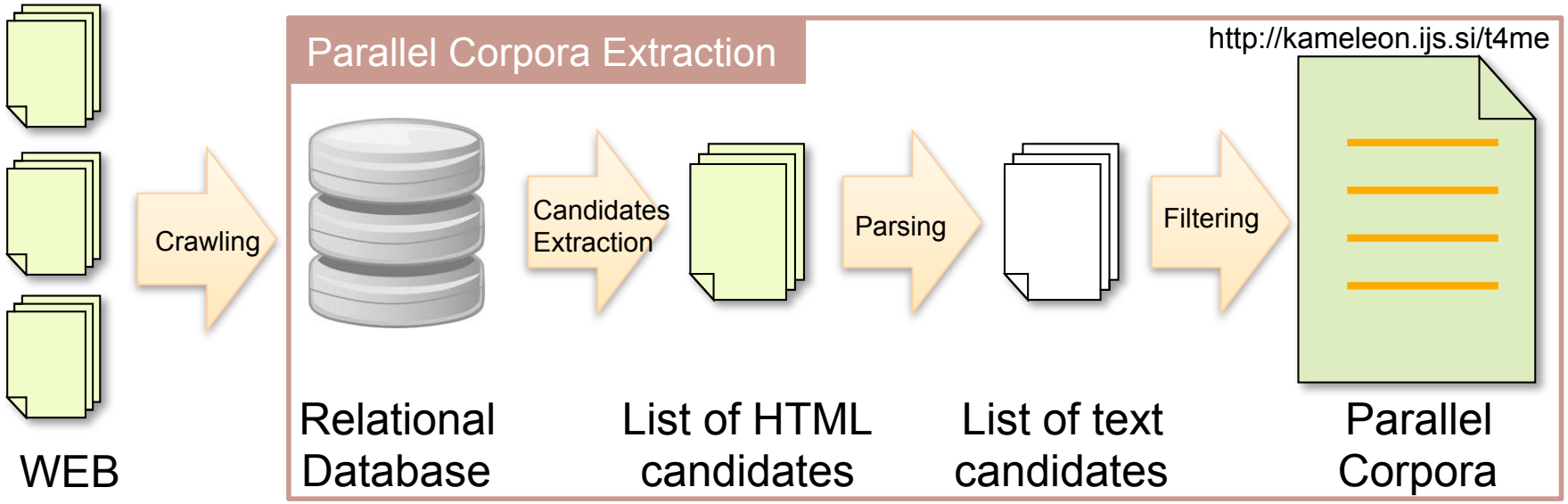


  - Clustering features:
    - same sentence, syntagmatic patterns, n-grams, collocational patterns

Data and Machine Learning for MT

Data and Machine Learning for MT

- There is no data like more data



- Cross-lingual document clustering (parallel text discovery)
  - Eurovoc keywords for training
  - SVD methods → (probabilistic) LSA for unsupervised, large data-based

Related Projects

Related 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

Future: the Challenges in 2012

- **Explore differences in architectures of MT Systems**
  - Task: to build Hybrid/System Combination systems
  - Data: Annotated Hybrid Sample MT Corpus (Spanish-to-English)
    - Provided by META-NET, train/test
  - Evaluation: Peer-based human evaluation
  - Venue: MT Summit XIII 2012, China (task period: May-July 2012)
- **Context in Translation Challenge (Preliminary)**
  - Task: Reranking Translation Candidates
  - Data: Eng/Fin, Gre/Fre, provided train/test
  - Evaluation: automatic (TBD)
  - Venue: ICANN 2012, Sept., Switzerland (task period: May-June 2012)

Q/A

Thank you very much.

office@meta-net.eu

http://www.meta-net.eu

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