ADAPTIVE INFORMATICS
RESEARCH CENTRE

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

Unsupervised Morpheme Analysis

Competition 3:

Statistical Machine Translation

Mikko Kurimo, Sami Virpioja, Ville T. Turunen (TKK)
Graeme W. Blackwood, William Byrne (UCAM)

Morphology and SMT

- Statistical machine translation systems find translation probabilities between words or sequences of words (“phrases”).
- Languages of rich morphology tend to be hard to translate both **from** and **to** – e.g. Finnish is one of the hardest among the EU languages.
- **Still unsolved problem**

Morph-based translation

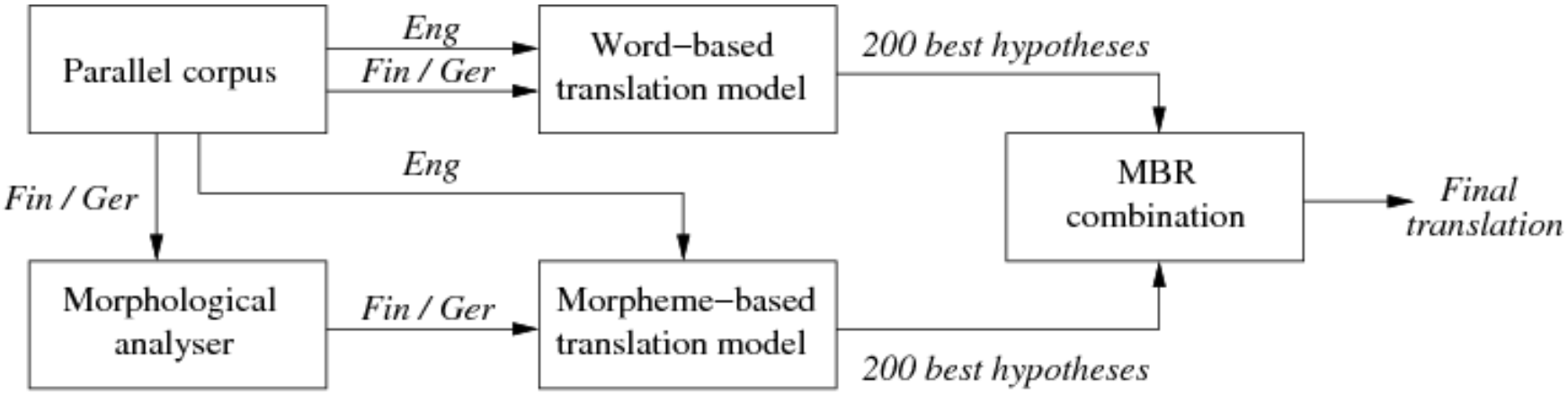
- Can unsupervised morphology learning directly improve SMT?
  - **Reduces out-of-vocabulary rates**
    (S. Virpioja, J. Väyrynen, M. Creutz & M. Sadeniemi, Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner, MT Summit XI, 2007)
  - **Improves translation results**
    (A. de Gispert, S. Virpioja, W. Byrne, M. Kurimo, Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions, HLT-NAACL, 2009)

Tasks and data

- Europarl parallel corpus
  - Proceedings of the EU parliament meetings in 11 European languages
- { **Finnish, German** } → English
  - Reducing OOV problems at the **source side**
  - Finnish: 479 780 word types
  - German: 270 038 word types
- ~1 million sentences for training, <3000 for tuning, 3000 for testing

System overview

- Evaluation based on combination of word-based and morph-based SMT systems (de Gispert et al., 2009)



Phrase-based SMT

- One of the major advances in SMT methodology in this decade
- Open source software: **Moses**
  (P. Koehn et al., 2007)
- Main steps in building a system with Moses:
  - Word alignment (Giza++)
  - Phrase extraction and scoring
  - Building additional models (language model, reordering model, etc.)
  - Parameter tuning for decoder

MBR and system combination

- **Minimum Bayes Risk (MBR) decoding:**
  - Select translation hypothesis which maximises the conditional expected gain:

$$\hat{E} = \underset{\hat{E} \in e}{argmax} \sum_{E \in e} G(E, \hat{E}) P(E|F)$$

- **System combination:** generate N-best lists from different systems and find the best hypothesis with the MBR criterion

MT evaluation

- There are several metrics for automatic evaluation of MT systems.
- **BLEU score** is based on co-occurrence of n-grams (n=1...4) in the proposed translation and the reference translation(s).
- Usually consistent with human evaluations if the evaluated systems are similar

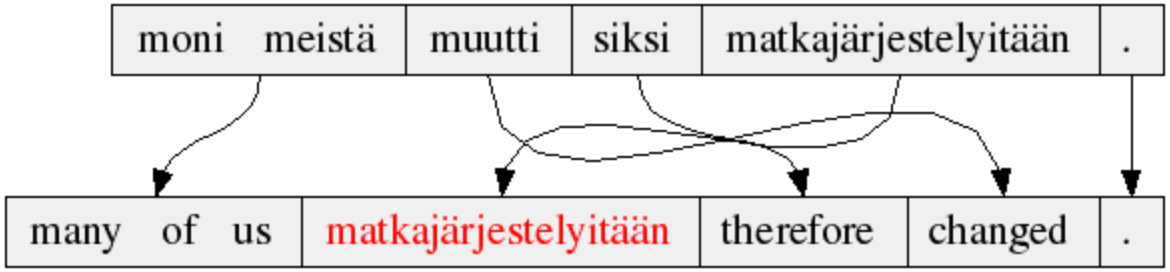
Submissions to Competition 3

- Bernhard – MorphoNet (MN)
- Monson et al. - ParaMor Mimic (PM)
- Monson et al. - ParaMor Morfessor Mimic (PMM)
- Monson et al. - ParaMor Morfessor Union (PMU)
- Virpioja & Kohonen – Allomorfessor (A)
- Tchoukalov et al. - MetaMorph (MM)

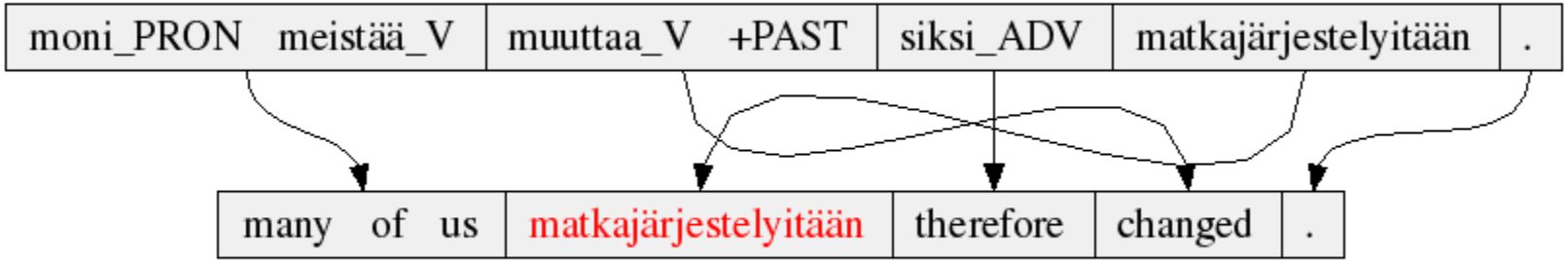
- Reference methods: Morfessor Baseline (MB), Morfessor CatMAP (MC), Grammatical (G)

Example translations (1)

Words

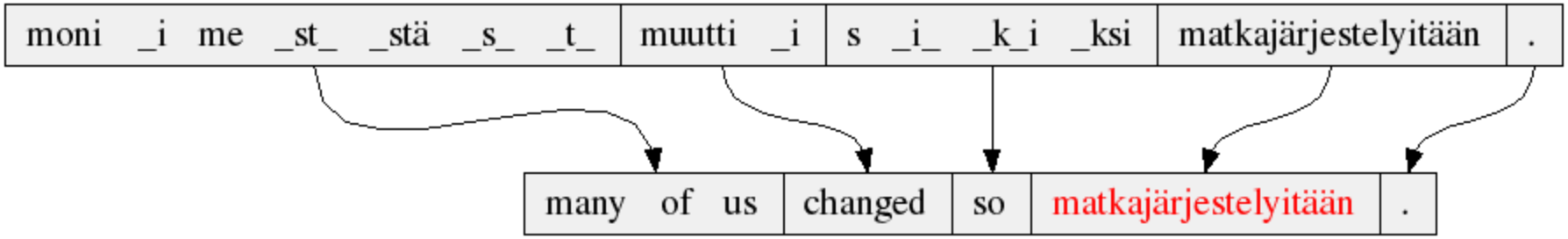


Grammatical gold standard

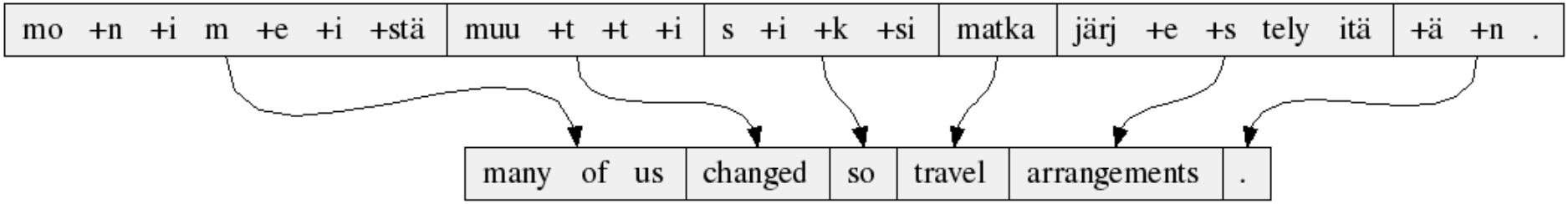


Example translations (2)

Bernhard - MorphoNet

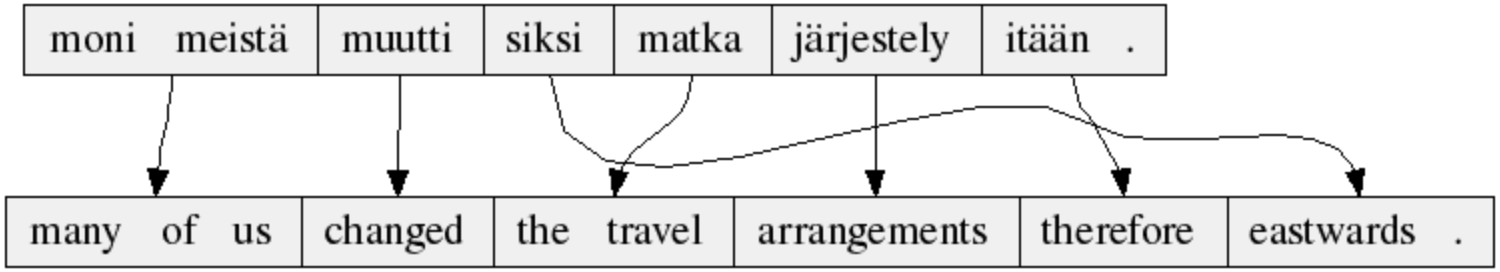


Monson et al. - ParaMor-Morfessor Union

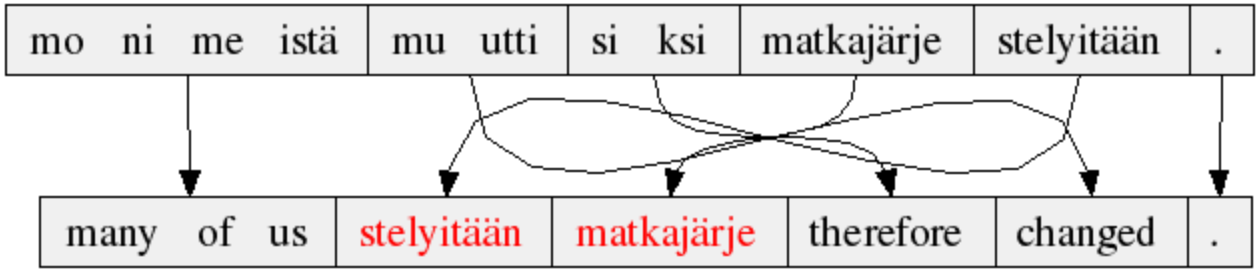


Example translations (3)

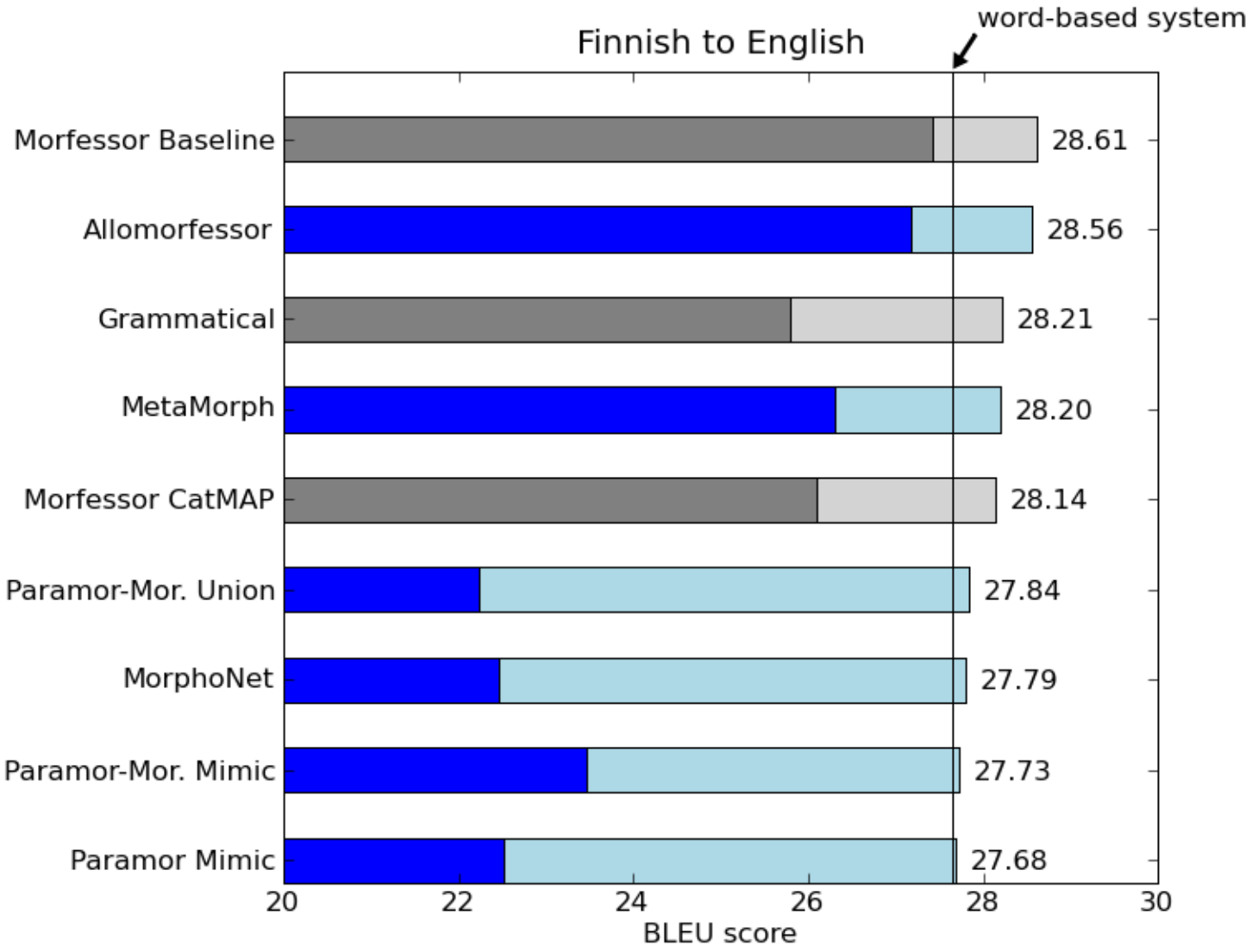
Virpioja & Kohonen - Allomorfessor



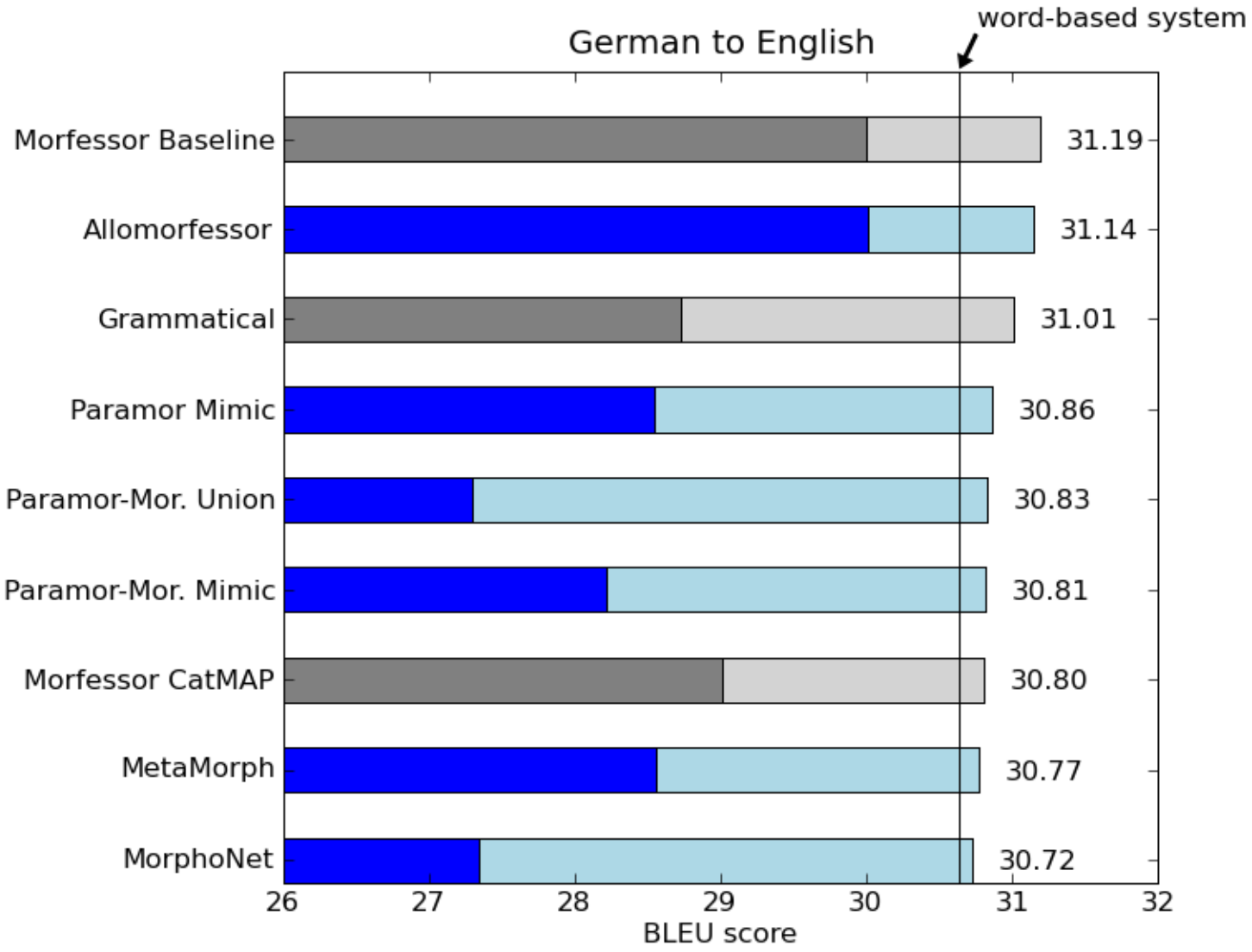
Tchoukalov et al. - MetaMorph



Results: Finnish



Results: German



Discussion

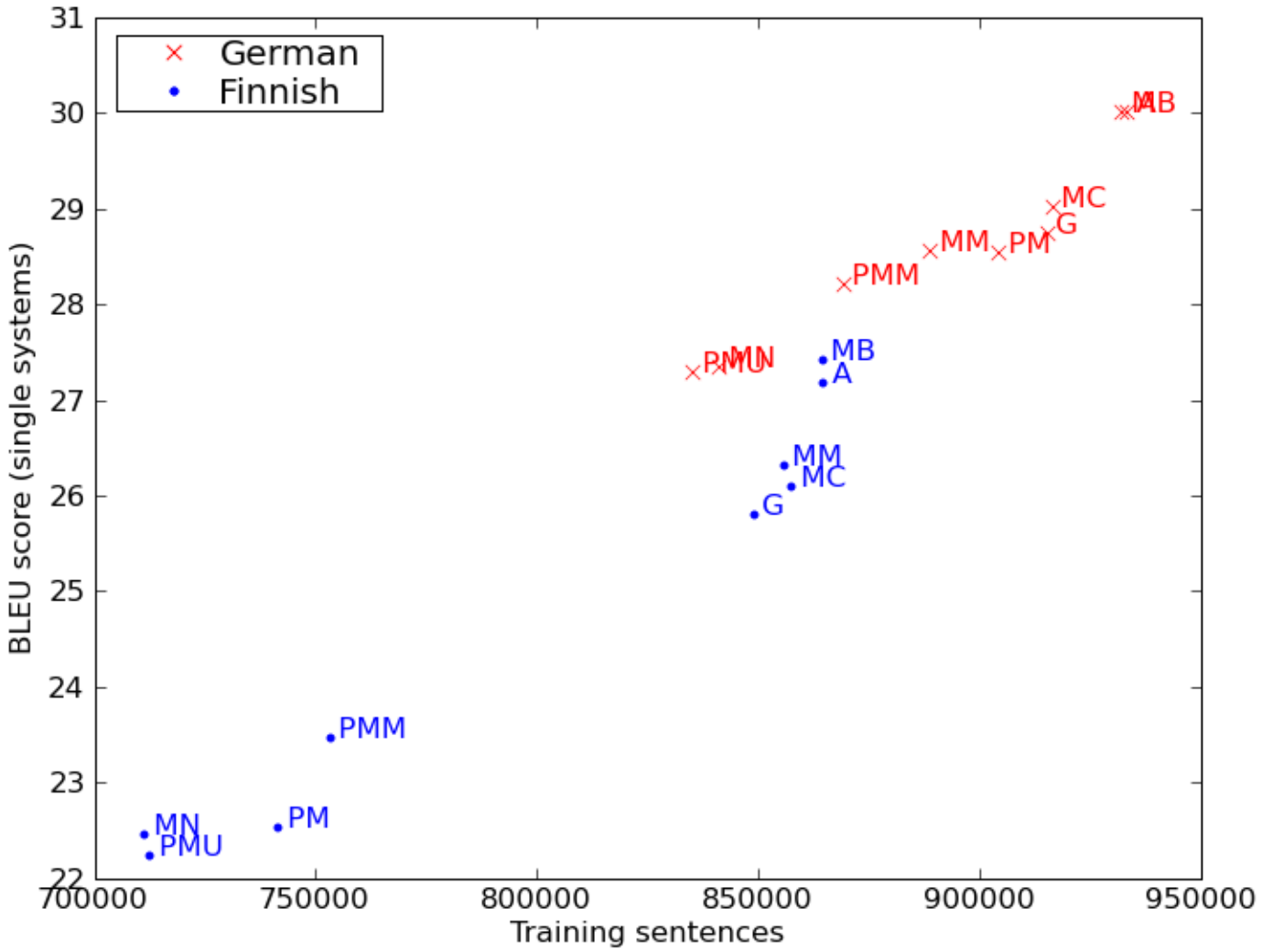
- Too long (>100 tokens) sentences cannot be handled by Giza++.
  - Segmentation decreases the amount of training data.
  - **Direct effect on performance**
- However, the number of average morphs per word does **not** explain the number of pruned sentences.

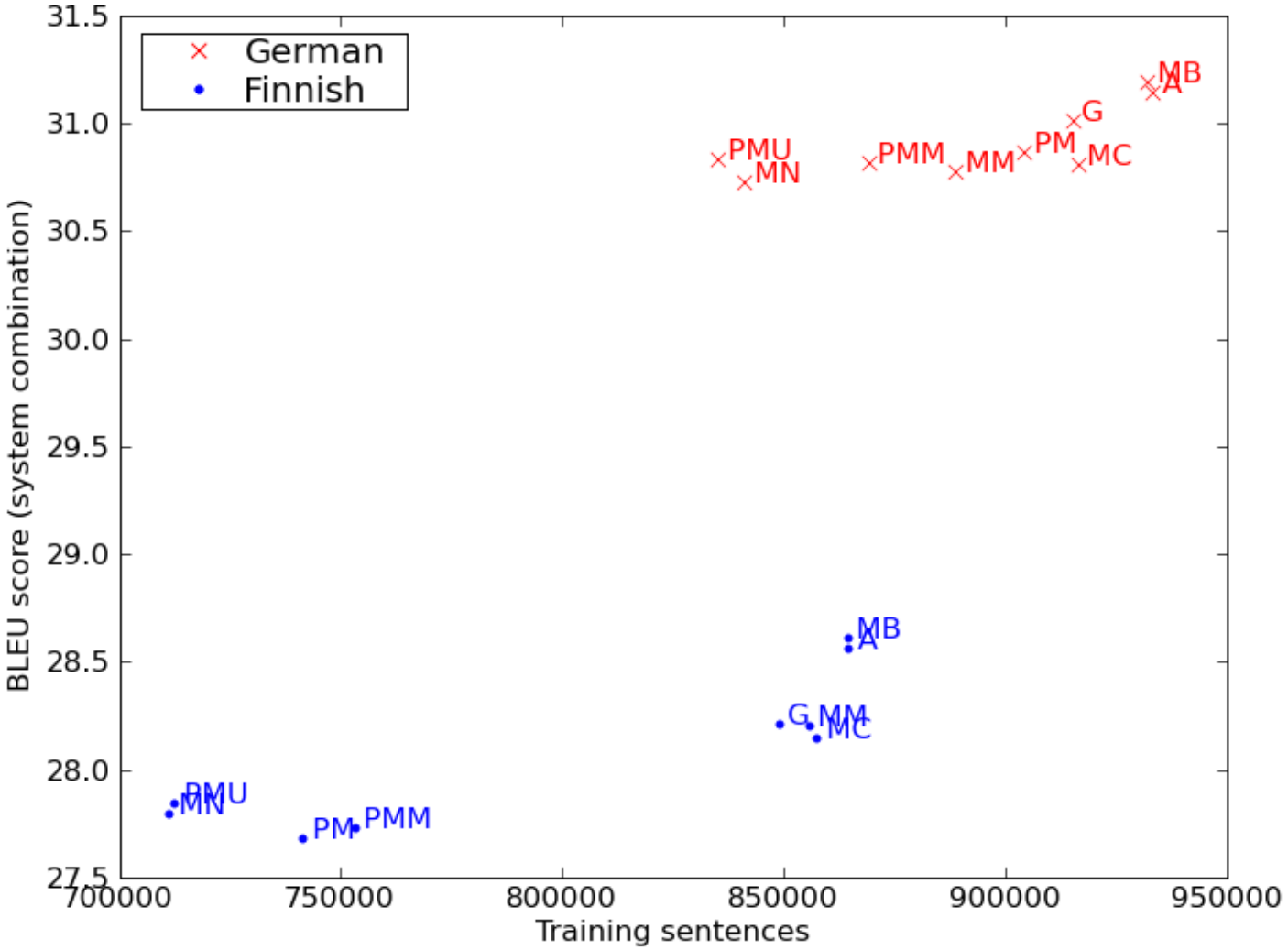
Conclusions

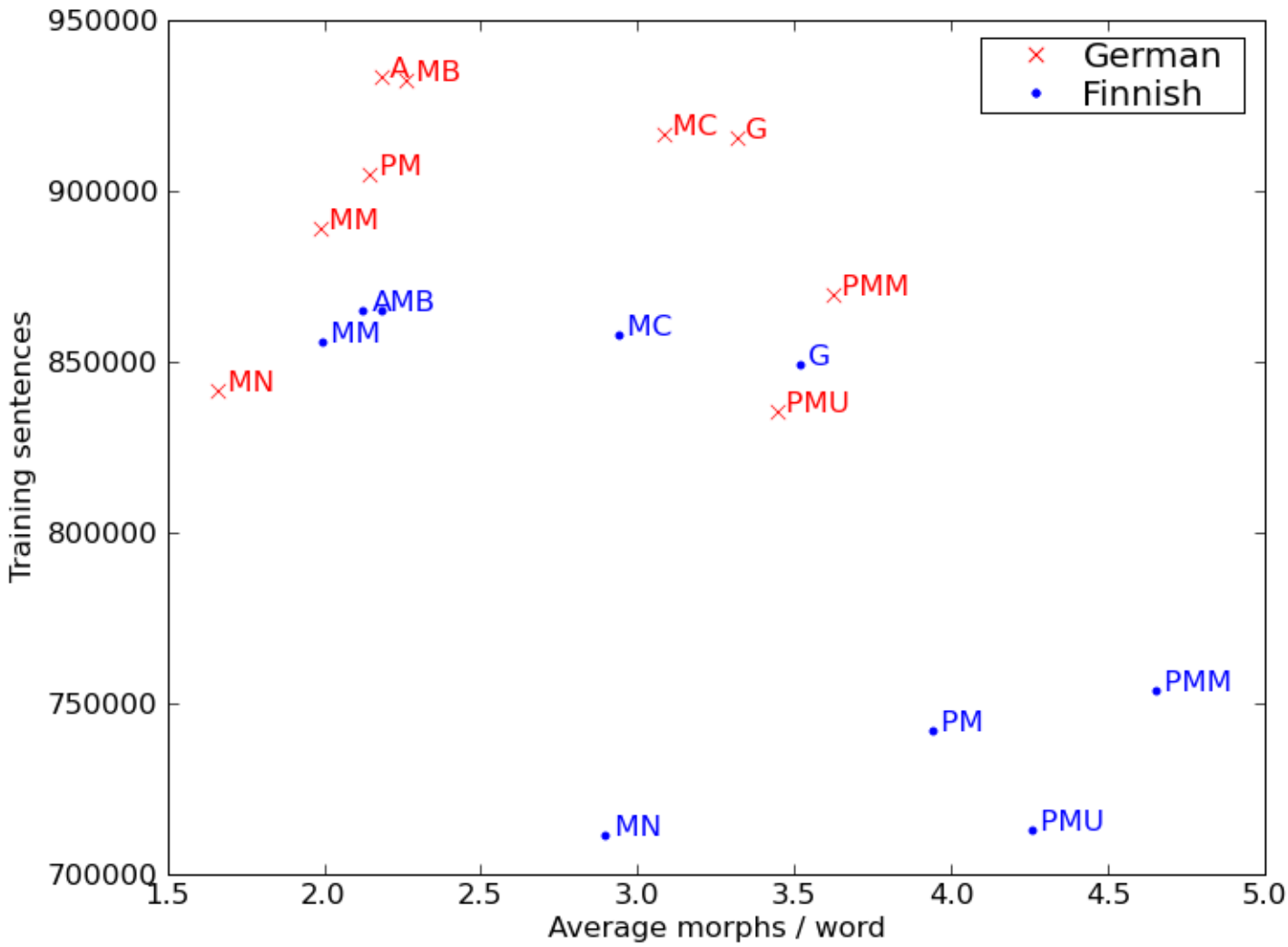
- 6 submitted and 3 reference methods were tested on two machine translation tasks.
- The 3-5 best methods improved the translation results over the baseline word-based system.
- Some improvements are needed to make the comparison more fair.
- Full report and papers in the CLEF proceedings
- Details, presentations, links, info at:

  http://www.cis.hut.fi/morphochallenge2009/

ADAPTIVE INFORMATICS
RESEARCH CENTRE


German
Finnish
BLEU score (single systems)
Training sentences
31
30
29
28
27
26
25
24
23
22
700000
750000
800000
850000
900000
950000
MB
MC
G
MM
PM
PMM
MN
PMU
A
PMM
MN
PM
PMU


German
Finnish
BLEU score (system combination)
Training sentences
31.5
31.0
30.5
30.0
29.5
29.0
28.5
28.0
27.5
700000
750000
800000
850000
900000
950000
MB
A
G
PMU
MN
PMM
MM
PM
MC
MB
A
G
MM
MC
PMU
MN
PM
PMM


German
Finnish
AMB
MC
G
PM
MM
PMM
AMB
MM
MC
G
MN
PMU
PMM
PM
MN
PMU
950000
900000
850000
800000
750000
700000
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Training sentences
Average morphs / word

MBR: A toy example

F = “Kahvi oli vahvaa.”

E1 = “The coffee was powerful.” P(E1 | F) = 0.4

E2 = “The coffee tasted strong.” P(E2 | F) = 0.4

E3 = “The coffee was strong.” P(E3 | F) = 0.2

G(x,y) = the number of common words

E1: 4 * 0.4 + 2 * 0.4 + 3 * 0.2 = 3.0

E2: 2 * 0.4 + 4 * 0.4 + 3 * 0.2 = 3.0

E3: 3 * 0.4 + 3 * 0.4 + 4 * 0.2 = **3.2**