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:

\[
\hat{E} = \arg\max_{E \in e} \sum_{\hat{E} \in e} G(E, \hat{E}) P(E|F)
\]

- **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**

<table>
<thead>
<tr>
<th>moni</th>
<th>meistä</th>
<th>muutti</th>
<th>siksi</th>
<th>matkajärjestelyitään</th>
</tr>
</thead>
</table>

| many of us | matkajärjestelyitään | therefore | changed |

**Grammatical gold standard**

<table>
<thead>
<tr>
<th>moni_PRON</th>
<th>meistää_V</th>
<th>muutaa_V</th>
<th>+PAST</th>
<th>siksi_ADV</th>
<th>matkajärjestelyitään</th>
</tr>
</thead>
</table>

| many of us | matkajärjestelyitään | therefore | changed |

.
Example translations (2)

Bernhard - MorphoNet

```
moni _i me __st_ __stä _s_ _t_ muutti _i s __k__i __ksi matkajärjestelyitään .
```

```
many of us changed so matkajärjestelyitään .
```

Monson et al. - ParaMor-Morfessor Union

```
mo +n +i m +e +i +stä muu +t +t +i s +i +k +si matka järj +e +s tely itä +ä +n .
```

```
many of us changed so travel arrangements .
```
Example translations (3)

Virpioja & Kohonen - Allomorfessor

moni meistä muutti siksi matka järjestyely itään .

many of us changed the travel arrangements therefore eastwards .

Tchoukalov et al. - MetaMorph

moni meistä muutti siksi matkajarje stelyitään .

many of us stelyitään matkajarje therefore changed .
Results: Finnish

- Morfessor Baseline: 28.61
- Allomorfessor: 28.56
- Grammatical: 28.21
- MetaMorph: 28.20
- Morfessor CatMAP: 28.14
- Paramor-Mor. Union: 27.84
- MorphoNet: 27.79
- Paramor-Mor. Mimic: 27.73
- Paramor Mimic: 27.68
Results: German

- Morfessor Baseline: 31.19
- Allomorfessor: 31.14
- Grammatical: 31.01
- Paramor Mimic: 30.86
- Paramor-Mor. Union: 30.83
- Paramor-Mor. Mimic: 30.81
- Morfessor CatMAP: 30.80
- MetaMorph: 30.77
- MorphoNet: 30.72

BLEU score range: 26 to 32
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:
MBR: A toy example

F = “Kahvi oli vahvaa.”
E1 = “The coffee was powerful.” \( P(E1 \mid F) = 0.4 \)
E2 = “The coffee tasted strong.” \( P(E2 \mid F) = 0.4 \)
E3 = “The coffee was strong.” \( P(E3 \mid F) = 0.2 \)

\[ G(x,y) = \text{the number of common words} \]

\begin{align*}
\text{E1:} & \quad 4 \times 0.4 + 2 \times 0.4 + 3 \times 0.2 = 3.0 \\
\text{E2:} & \quad 2 \times 0.4 + 4 \times 0.4 + 3 \times 0.2 = 3.0 \\
\text{E3:} & \quad 3 \times 0.4 + 3 \times 0.4 + 4 \times 0.2 = 3.2
\end{align*}