Aalto University

MT at Aalto: Data-driven morphological segmentation for SMT

Stig-Arne Grönroos and Sami Virpioja

Aalto University, Finland
stig-arne.gronroos@aalto.fi
sami.virpioja@aalto.fi

12.09.2016

MT at Aalto University

Who?

- Mathias Creutz, Marcus Dobrinkat, Stig-Arne Grönroos, Timo Honkela, Mikko Kurimo, André Mansikkaniemi, Markus Sadeniemi, Tero Tapiovaara, Sami Virpioja, Jaakko Väyrynen

MT at Aalto University

Who?

- Mathias Creutz, Marcus Dobrinkat, Stig-Arne Grönroos, Timo Honkela, Mikko Kurimo, André Mansikkaniemi, Markus Sadeniemi, Tero Tapiovaara, Sami Virpioja, Jaakko Väyrynen

Where (and when)?

- Lab. of Computer and Information Science (2006–2007),
  Dept. of Information and Computer Science (2008–2014),
  Dept. of Computer Science (2015–2016),
  Dept. of Signal Processing and Acoustics (2013–)

MT at Aalto University

Who?

- Mathias Creutz, Marcus Dobrinkat, Stig-Arne Grönroos, Timo Honkela, Mikko Kurimo, André Mansikkaniemi, Markus Sadeniemi, Tero Tapiovaara, Sami Virpioja, Jaakko Väyrynen

Where (and when)?

- Lab. of Computer and Information Science (2006–2007),
  Dept. of Information and Computer Science (2008–2014),
  Dept. of Computer Science (2015–2016),
  Dept. of Signal Processing and Acoustics (2013–)

What?

- Morphological segmentation for SMT
- Machine translation evaluation
  - mNCD (Dobrinkat et al., 2010a,b): Normalized compression distance as MT metric
  - LeBLEU (Virpioja and Grönroos, 2015): Variant of BLEU with scoring based on letter-edit distances

MT at Aalto University

Who?

- Mathias Creutz, Marcus Dobrinkat, Stig-Arne Grönroos, Timo Honkela, Mikko Kurimo, André Mansikkaniemi, Markus Sadeniemi, Tero Tapiovaara, Sami Virpioja, Jaakko Väyrynen

Where (and when)?

- Lab. of Computer and Information Science (2006–2007),
  Dept. of Information and Computer Science (2008–2014),
  Dept. of Computer Science (2015–2016),
  Dept. of Signal Processing and Acoustics (2013–)

What?

- Morphological segmentation for SMT
- Machine translation evaluation
  - mNCD (Dobrinkat et al., 2010a,b): Normalized compression distance as MT metric
  - LeBLEU (Virpioja and Grönroos, 2015): Variant of BLEU with scoring based on letter-edit distances

Unsupervised segmentation for SMT

Motivation:

- Unsuperived segmentation useful for reducing lexicon size in Finnish ASR (Hirsimäki et al., 2006)
- Morfessor Categories-MAP (Creutz and Lagus, 2007) suitable for SMT
  - $\text{erä}^{+}_{\text{STM}}$ itä_{SUF} $\text{sääntely}^{+}_{\text{STM}}$ $\text{toimi}^{+}_{\text{STM}}$ a_{SUF} on_{STM} ...

Unsupervised segmentation for SMT

Motivation:

- Unsuperived segmentation useful for reducing lexicon size in Finnish ASR (Hirsimäki et al., 2006)
- Morfessor Categories-MAP (Creutz and Lagus, 2007) suitable for SMT
  - $\text{erä}^{+}_{\text{STM}}$ itä_{SUF} $\text{sääntely}^{+}_{\text{STM}}$ $\text{toimi}^{+}_{\text{STM}}$ a_{SUF} on_{STM} ...

Method:

- Segment both source and target with Morfessor Categories-MAP before training a phrase-based translation model.
- Can be applied to (almost) any language pair.

Unsupervised segmentation for SMT

Motivation:

- Unsuperived segmentation useful for reducing lexicon size in Finnish ASR (Hirsimäki et al., 2006)
- Morfessor Categories-MAP (Creutz and Lagus, 2007) suitable for SMT
  - $\text{erä}^{+}_{\text{STM}}$ itä_{SUF} $\text{sääntely}^{+}_{\text{STM}}$ $\text{toimi}^{+}_{\text{STM}}$ a_{SUF} on_{STM} ...

Method:

- Segment both source and target with Morfessor Categories-MAP before training a phrase-based translation model.
- Can be applied to (almost) any language pair.

Results:

- Virpioja et al. (2007): da $\rightleftarrows$ fi, da $\rightleftarrows$ sv, fi $\rightleftarrows$ sv
  - No improvement in BLEU
  - Reduced out-of-vocabulary rate
  - Some useful phrases that end or start with morphs

Unsupervised segmentation for SMT

Motivation:

- Unsuperived segmentation useful for reducing lexicon size in Finnish ASR (Hirsimäki et al., 2006)
- Morfessor Categories-MAP (Creutz and Lagus, 2007) suitable for SMT
  - $\text{erä}^{+}_{\text{STM}}$ itä_{SUF} $\text{sääntely}^{+}_{\text{STM}}$ $\text{toimi}^{+}_{\text{STM}}$ a_{SUF} on_{STM} ...

Method:

- Segment both source and target with Morfessor Categories-MAP before training a phrase-based translation model.
- Can be applied to (almost) any language pair.

Results:

- Virpioja et al. (2007): da $\rightleftarrows$ fi, da $\rightleftarrows$ sv, fi $\rightleftarrows$ sv
  - No improvement in BLEU
  - Reduced out-of-vocabulary rate
  - Some useful phrases that end or start with morphs

Next question:

- How to make use of the reduced OOV rate without degrading the overall result?

Combining alternative decompositions

Motivation:

- If target language is not split, several segmentation approaches for source language can be combined.

Combining alternative decompositions

Motivation:

- If target language is not split, several segmentation approaches for source language can be combined.

Method:

- Segment complex source language with Morfessor Categories-MAP.
- Perform Minimum Bayes Risk combination with a word-based translation model.

Combining alternative decompositions

Motivation:

- If target language is not split, several segmentation approaches for source language can be combined.

Method:

- Segment complex source language with Morfessor Categories-MAP.
- Perform Minimum Bayes Risk combination with a word-based translation model.

Results:

- de Gispert et al. (2009): Improves BLEU for fi $\rightarrow$ en

Combining alternative decompositions

Motivation:

- If target language is not split, several segmentation approaches for source language can be combined.

Method:

- Segment complex source language with Morfessor Categories-MAP.
- Perform Minimum Bayes Risk combination with a word-based translation model.

Results:

- de Gispert et al. (2009): Improves BLEU for fi $\rightarrow$ en

Next question:

- Morphological segmentation increases number of tokens per sentence, making alignment and translation more complex. How to prevent oversegmentation?

Weighted Morfessor Baseline segmentation

Motivation:

- The segmentation of Morfessor Baseline can be easily tuned by modifying the frequencies of input words (Virpioja et al., 2011a).
- High precision of morphological segmentation was much more important than recall in SMT evaluation of Morpho Challenges 2009–2010 (Virpioja et al., 2011b).

Weighted Morfessor Baseline segmentation

Motivation:

- The segmentation of Morfessor Baseline can be easily tuned by modifying the frequencies of input words (Virpioja et al., 2011a).
- High precision of morphological segmentation was much more important than recall in SMT evaluation of Morpho Challenges 2009–2010 (Virpioja et al., 2011b).

Method:

- Control segmentation of the complex source language using frequency-weighted Morfessor Baseline.
- Apply MBR combination with word-based model.

Weighted Morfessor Baseline segmentation

Motivation:

- The segmentation of Morfessor Baseline can be easily tuned by modifying the frequencies of input words (Virpioja et al., 2011a).
- High precision of morphological segmentation was much more important than recall in SMT evaluation of Morpho Challenges 2009–2010 (Virpioja et al., 2011b).

Method:

- Control segmentation of the complex source language using frequency-weighted Morfessor Baseline.
- Apply MBR combination with word-based model.

Results:

- Virpioja et al. (2010): Weighted Morfessor Baseline better than Categories-MAP in de $\rightarrow$ en

Weighted Morfessor Baseline segmentation

Motivation:

- The segmentation of Morfessor Baseline can be easily tuned by modifying the frequencies of input words (Virpioja et al., 2011a).
- High precision of morphological segmentation was much more important than recall in SMT evaluation of Morpho Challenges 2009–2010 (Virpioja et al., 2011b).

Method:

- Control segmentation of the complex source language using frequency-weighted Morfessor Baseline.
- Apply MBR combination with word-based model.

Results:

- Virpioja et al. (2010): Weighted Morfessor Baseline better than Categories-MAP in de $\rightarrow$ en

Next question:

- Morfessor Baseline is a bit too crude for SMT. Can we tune a Categories-MAP model?

Tuning Morfessor FlatCat for SMT

Motivation:

- Morfessor FlatCat (Grönroos et al., 2014):
  - Category-based model similar to Categories-MAP
  - Possible to tune the granularity of segmentation

Tuning Morfessor FlatCat for SMT

Motivation:

- Morfessor FlatCat (Grönroos et al., 2014):
  - Category-based model similar to Categories-MAP
  - Possible to tune the granularity of segmentation

Method:

- Tune the segmentation of morphologically complex language to produce similar amount of tokens per sentence as in the less complex language.
- Test different boundary markings for composition of segmented words.
- Rescore with RNNLM language model.

Tuning Morfessor FlatCat for SMT

Motivation:

- Morfessor FlatCat (Grönroos et al., 2014):
  - Category-based model similar to Categories-MAP
  - Possible to tune the granularity of segmentation

Method:

- Tune the segmentation of morphologically complex language to produce similar amount of tokens per sentence as in the less complex language.
- Test different boundary markings for composition of segmented words.
- Rescore with RNNLM language model.

Results:

- Grönroos et al. (2015): Significant improvement for en $\rightarrow$ fi out-of-domain translation (without system combination!)

Tuning Morfessor FlatCat for SMT

Motivation:

- Morfessor FlatCat (Grönroos et al., 2014):
  - Category-based model similar to Categories-MAP
  - Possible to tune the granularity of segmentation

Method:

- Tune the segmentation of morphologically complex language to produce similar amount of tokens per sentence as in the less complex language.
- Test different boundary markings for composition of segmented words.
- Rescore with RNNLM language model.

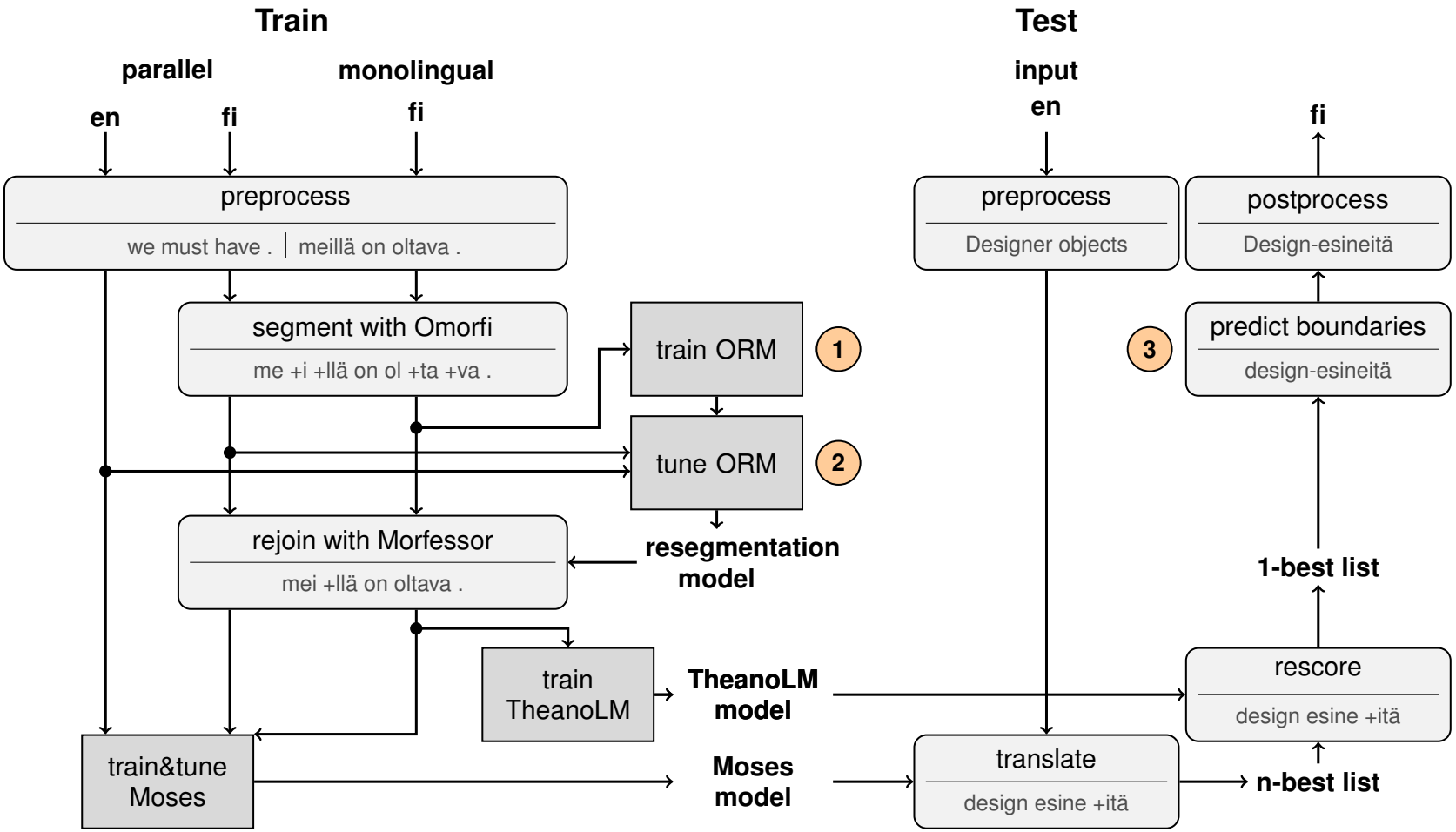
Results:

- Grönroos et al. (2015): Significant improvement for en $\rightarrow$ fi out-of-domain translation (without system combination!)

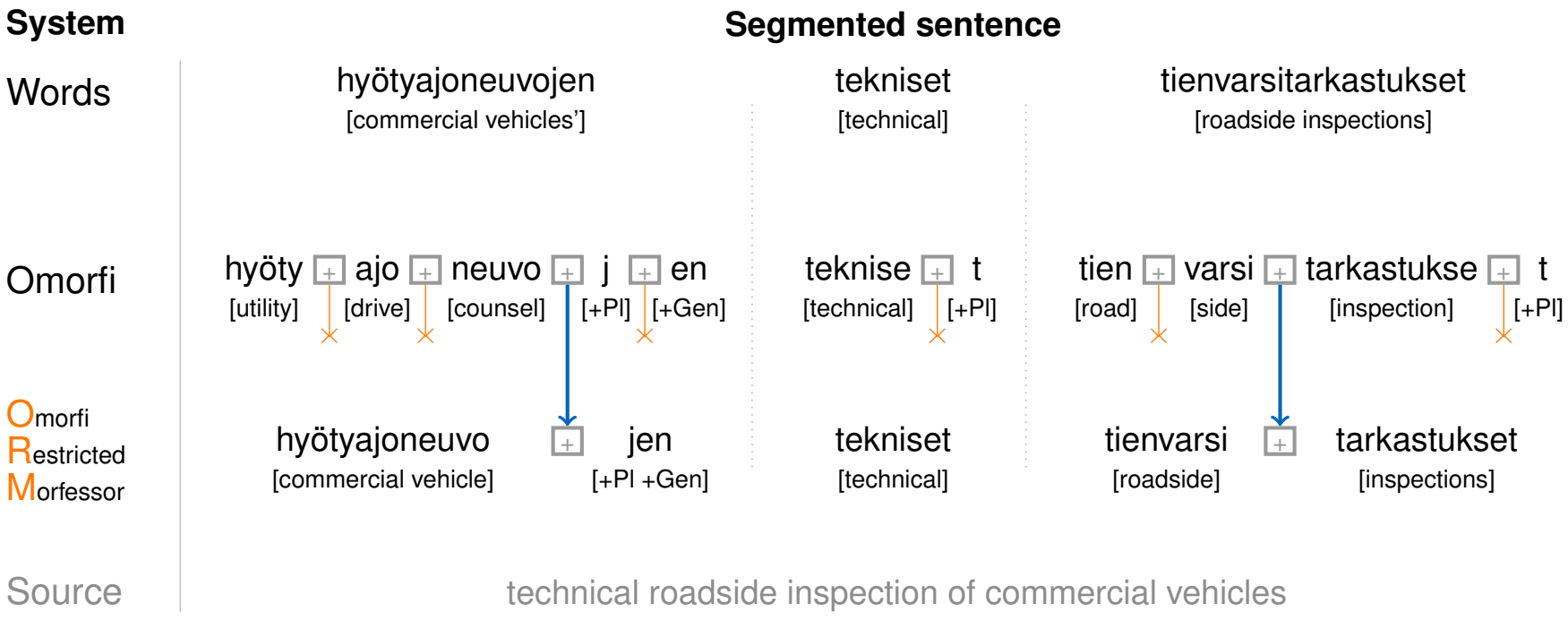
Next question:

- We have tools for rule-based morphological segmentation. Can we automatically tune their output for SMT?

Grönroos et al. (2016), System overview



1 Omorfi-Restricted Morfessor



2 Tuning the segmentation

Bigger lexicon, less segmentation –  – Smaller lexicon, more segmentation

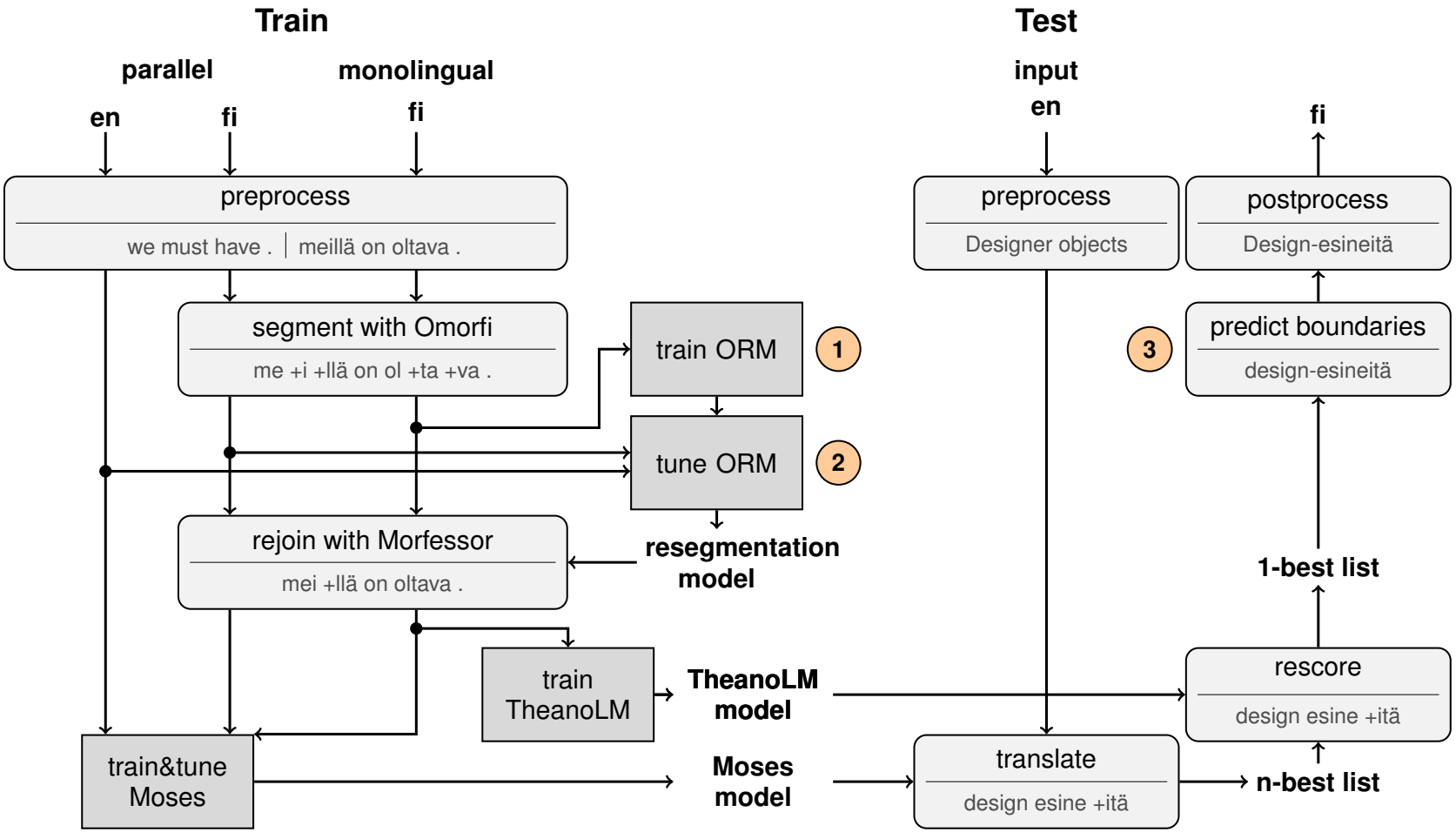
$$L(D,\theta) = L(\theta) + \alpha L(D \mid \theta)$$

$L(\theta)$: Morph lexicon

$L(D \mid \theta)$: Corpus encoded w/ lexicon

- Tune the segmentation of morphologically complex language to produce similar amount of tokens per sentence as in the less complex language

Grönroos et al. (2016), System overview



3 Correcting morph boundary markers

moni + liberaalien keskuudessa

[multi-] [liberals] [among]

[among the multiliberals]

3 Correcting morph boundary markers

moni	liberaalien	keskuudessa
[many]	[liberals]	[among]

~~[among the multiliberals]~~

[many among the liberals]

3 Correcting morph boundary markers

moni liberaalien keskuudessa

[many] [liberals] [among]

~~[among the multiliberals]~~

[many among the liberals]

opinto ⊞ opas

[study] [guide]

3 Correcting morph boundary markers

moni liberaalien keskuudessa

[many] [liberals] [among]

~~[among the multiliberals]~~

[many among the liberals]

opinto - opas

[study] [guide]

Future work

- Word type -level tuning of Morfessor
  - Using adjusted token counts
- More powerful post-prediction
  - Stymne and Cancedda (2011); Cap et al. (2014)
- Applying to NMT

Summary

Development of Morfessor for SMT:

1. Find efficient segmentation for your monolingual corpus
2. Tune segmentation for your parallel corpus
3. Restrict segmentation with your linguistic gold standard

Open source software:

- Morfessor Baseline
  - `https://github.com/aalto-speech/morfessor`
- Restricted Morfessor (ORM) and tuning scripts
  - `feat_typelevel_alignedtokencount` branch
- Morfessor FlatCat
  - `https://github.com/aalto-speech/flatcat`

Bibliography

Fabienne Cap, Alexander M Fraser, Marion Weller, and Aoife Cahill. How to produce unseen teddy bears: Improved morphological processing of compounds in SMT. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34, January 2007.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73–76, Boulder, USA, June 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N/N09/N09-2019.pdf`.

Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. Evaluating machine translations using mNCD. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 80–85. Association for Computational Linguistics, 2010a.

Bibliography (cont.)

Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. Normalized compression distance based measures for MetricsMATR 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 343–348. Association for Computational Linguistics, 2010b.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. ISBN 978-1-941643-26-6. URL `http://www.aclweb.org/anthology/C14-1111`.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. Tuning phrase-based segmented translation for a morphologically complex target language. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 105–111, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-3010`.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. Hybrid morphological segmentation for phrase-based machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 289–295, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W16/W16-2312.pdf`.

Bibliography (cont.)

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541, October 2006. URL `http://www.sciencedirect.com/science/article/B6WCW-4H09XJ2-1/2/4a495d6b45f046d9d0b74389b4914689`.

Sara Stymne and Nicola Cancedda. Productive generation of compound words in statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, 2011. Association for Computational Linguistics.

Sami Virpioja and Stig-Arne Grönroos. LeBLEU: N-gram-based translation evaluation score for morphologically complex languages. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 411–416, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-3052`.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark, September 2007.

Bibliography (cont.)

Sami Virpioja, André Mansikkaniemi, Jaakko Väyrynen, and Mikko Kurimo. Applying morphological decompositions to statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206. Association for Computational Linguistics, July 2010. URL `http://www.statmt.org/wmt10/pdf/WMT30.pdf`.

Sami Virpioja, Oskar Kohonen, and Krista Lagus. Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In Bolette Sandford Pedersen, Gunta Nešpore, and Inguna Skadiņa, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of *NEALT Proceedings Series*, pages 230–237. Northern European Association for Language Technology, Riga, Latvia, May 2011a. URL `http://hdl.handle.net/10062/17313`.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90, 2011b. URL `http://www.atala.org/Empirical-Comparison-of-Evaluation`.