

Proposals for master theses

Academic year 2017-18

Antonio Toral

`https://antoniotor.al`

`a.toral.ruiz@rug.nl`

University of Groningen

December 13, 2017

Proposals

1. Word Segmentation in Neural MT
2. Cross-sentence context for translating novels

Problem

Aim: better word segmentation for Morphologically Rich Languages (MPL), e.g. Finnish, Turkish.

Problem

Aim: better word segmentation for Morphologically Rich Languages (MPL), e.g. Finnish, Turkish.

Current situation

- ▶ No segmentation: severe vocabulary limit, <100K most frequent words
- ▶ Byte-pair encoding (BPE) [Sennrich et al., 2015]: unsupervised & good results, but not for MPL
- ▶ Morph segmentation [Sánchez-Cartagena and Toral, 2016]: good results for MPL, but supervised & leads to long sequences

Task

Segmentation	Sentence
None	haluaisimme , että oppisimme tästä yhden perusasian
BPE: 60k ops	haluaisimme , että opp→ ←-isimme tästä yhden perusasi→ ←-an
Omorfi	halua→ ←-isi→ ←-mme , että opp→ ←-isi→ ←-mme tästä yhde→ ←-n perus→ ←-asia→ ←-n
Omorfi + BPE: 1k ops.	halua→ ←-isimme , että opp→ ←-isimme tästä yhden perus→ ←-asian
English	<i>there is one basic lesson I would like us to learn from this</i>

Example of the application of the different segmentation schemes to a Finnish sentence.

- ▶ Note the compound word *perusasian*: *perus* (“basic”) + *asia* (“thing, affair”) + case marker *-n*
- ▶ How is the compound word *perusasian* segmented by the different schemes?

Task

1. Improve word segmentation for MRLs. For example:
 - ▶ Explore unsupervised morph segmentation
 - ▶ Explore in detail the joint use of morph segmentation + BPE
 - ▶ Use multiple segmentations in parallel, e.g. lattice
 - ▶ ...
2. Test in a shared task: WMT 2018 (Finnish, Turkish, Estonian)

What is there

- ▶ Segmenters: BPE, Omorfi, morfessor
- ▶ Strong baseline system for English–Finnish (winner WMT'16) [Sánchez-Cartagena and Toral, 2016]

Plan

Word Segmentation in Neural MT

Cross-sentence context for translating novels

Problem

Aim: model cross-sentence context for translation of novels

Current situation

- ▶ Sentences translated in isolation in MT
- ▶ Literary texts have dense reference chains → context beyond sentence essential for translation [Voigt and Jurafsky, 2012]
- ▶ First approaches to model cross-sentence context in neural MT [Wang et al., 2017]





Task

- ▶ Investigate and improve the neural MT system with context
- ▶ Build two neural MT systems tailored to novels for e.g. English→Dutch, with and without cross-sentence context
- ▶ Evaluate the impact on the translations

What is there

- ▶ Neural MT toolkits
- ▶ Experimental neural MT system with cross-sentence context
- ▶ Neural MT system tailored to novels
- ▶ GUI for fine-grained linguistic evaluation

References I

-  Sánchez-Cartagena, V. M. and Toral, A. (2016).
Abu-matran at wmt 2016 translation task: Deep learning,
morphological segmentation and tuning on character sequences.
-  Sennrich, R., Haddow, B., and Birch, A. (2015).
Neural machine translation of rare words with subword units.
arXiv preprint arXiv:1508.07909.
-  Voigt, R. S. U. and Jurafsky, D. S. U. (2012).
Towards a Literary Machine Translation: The Role of Referential
Cohesion.
*The 2012 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language
Technologies (NAACL-HLT 2012)*, pages 18–25.
-  Wang, L., Tu, Z., Way, A., and Liu, Q. (2017).
Exploiting cross-sentence context for neural machine translation.
CoRR, abs/1704.04347.

Thank you!

Questions?