Aalto University

North Sámi morphological segmentation with low-resource semi-supervised sequence labeling

Stig-Arne Grönroos, Sami Virpioja and Mikko Kurimo

Department of Signal Processing and Acoustics
Aalto University, School of Electrical Engineering
stig-arne.gronroos@aalto.fi

7.1.2019

Introduction

From the CFP:

- Are trendy deep-learning methods applicable to Uralic languages?
  - Challenge: Low resource
  - Challenge: Rich morphology
- Comparative evaluation of different open-source NLP methods as applied to Uralic languages.
  - Task: morphological surface segmentation
  - Language: North Sámi

Seq2seq – a versatile NLP tool

- Machine translation
  - (Vaswani et al., 2017) 4.5M/36M sent pairs
- Summarization
  - (Nallapati et al., 2016) 3.8M examples
- Speech synthesis
  - (Wang et al., 2017) 24.6h speech
- . . .

Seq2seq – a versatile NLP tool

- Numerous tasks can be described as mapping from sequence to sequence.

$$x \mapsto y$$

- Including several morphological processing tasks.

1. Choose an NMT system
2. Preprocess your data
3. Profit

Morphological tasks: Morphological analysis

- Yields the lemma and tags.

e.g. *took* $\mapsto$ *take* PAST

$w \mapsto yt; \quad w, y \in \Sigma^*, t \in \tau^*$

Morphological tasks: Morphological analysis

- Yields the lemma and tags.

e.g. *took* $\mapsto$ *take* PAST

$w \mapsto yt; \quad w, y \in \Sigma^*, t \in \tau^*$

Morphological tasks: Reinflection

- Yields another inflected form.

e.g. *taken* PAST $\mapsto$ *took*

$wt \mapsto y; \quad w, y \in \Sigma^*, t \in \tau^*$

Morphological tasks: Lemmatization

- Yields the lemma.

$$\text{e.g.} \quad \textit{better} \mapsto \textit{good}$$

$$w \mapsto y; \quad w, y \in \Sigma^*$$

Morphological tasks: Canonical segmentation

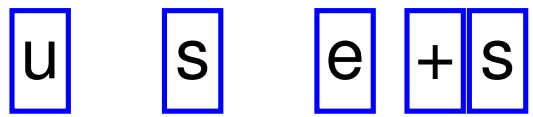
- Yields standardized segments.
- Undo morphological processes.

e.g. $\textit{achievability} \mapsto \textit{achieve} + \textit{able} + \textit{ity}$

$$w \mapsto y; \quad w \in \Sigma^*, y \in (\Sigma \cup \{+\})^*$$

Morphological tasks: Surface segmentation

$$w \mapsto y; \quad w \in \Sigma^*, y \in (\Sigma \cup \{+\})^*$$

uses $\mapsto$ 

Morphological tasks: Surface segmentation

$$w \mapsto y; \quad w \in \Sigma^*, y \in (\Sigma \cup \{+\})^*$$

- Yields morphs that concatenate back to the word.
- No need to generate arbitrary length sequence.

0 u 0 s 0 e 1 s

Morphological tasks: Surface segmentation

~~$w \mapsto y; \quad w \in \Sigma^*, y \in (\Sigma \cup \{+\})^*$~~

- Yields morphs that concatenate back to the word.
- No need to generate arbitrary length sequence.

$$w \mapsto y; \quad w \in \Sigma^k, y \in \Omega^k, k \in \mathbb{N}$$

out:	0	0	0	1
in:	u	s	e	s

Morphological tasks: Surface segmentation

~~$w \mapsto y; \quad w \in \Sigma^*, y \in (\Sigma \cup \{+\})^*$~~

- Yields morphs that concatenate back to the word.
- No need to generate arbitrary length sequence.

$$w \mapsto y; \quad w \in \Sigma^k, y \in \Omega^k, k \in \mathbb{N}$$

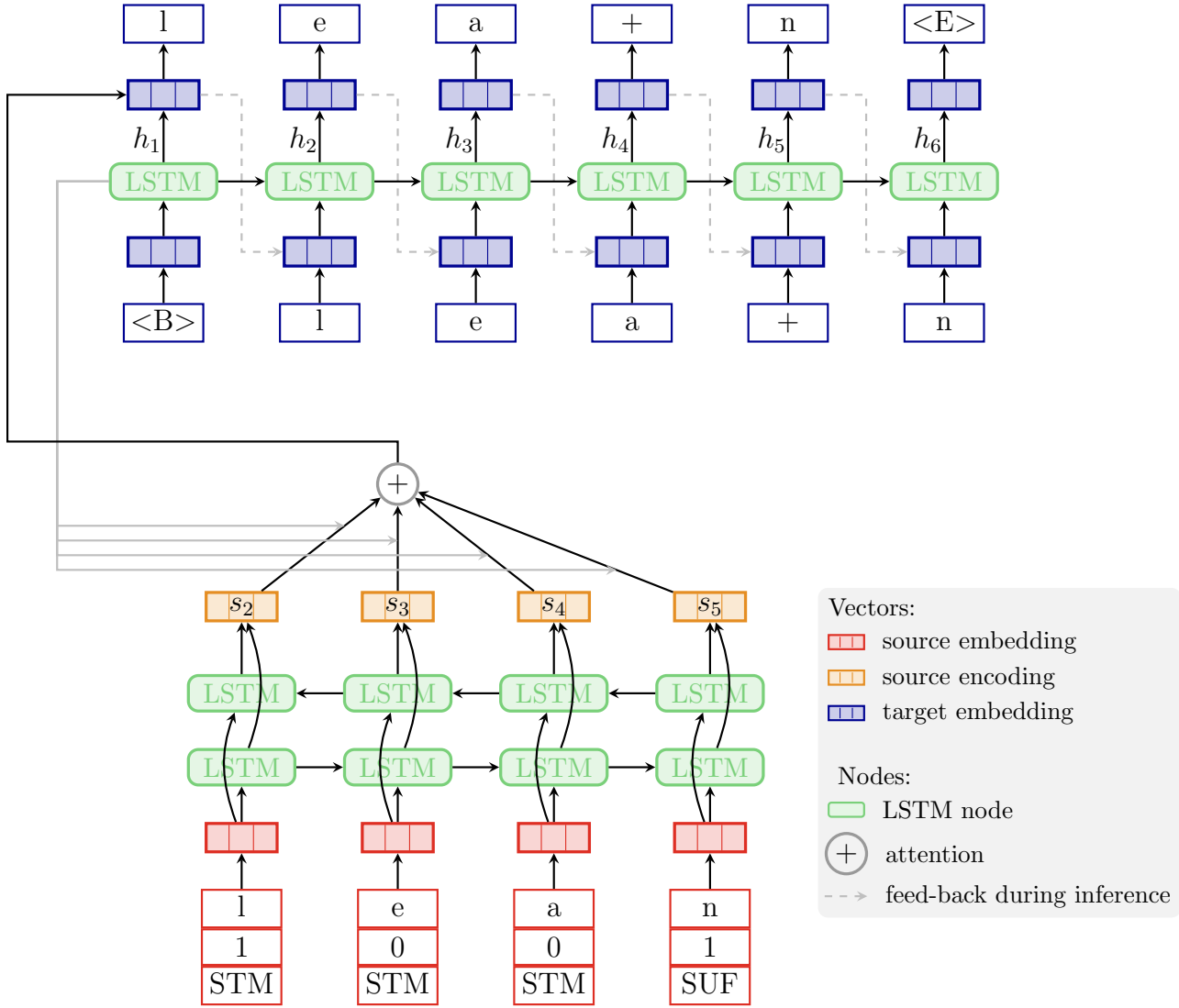
e.g. $uses \mapsto BMES$

out:	B	M	E	S
in:	u	s	e	s

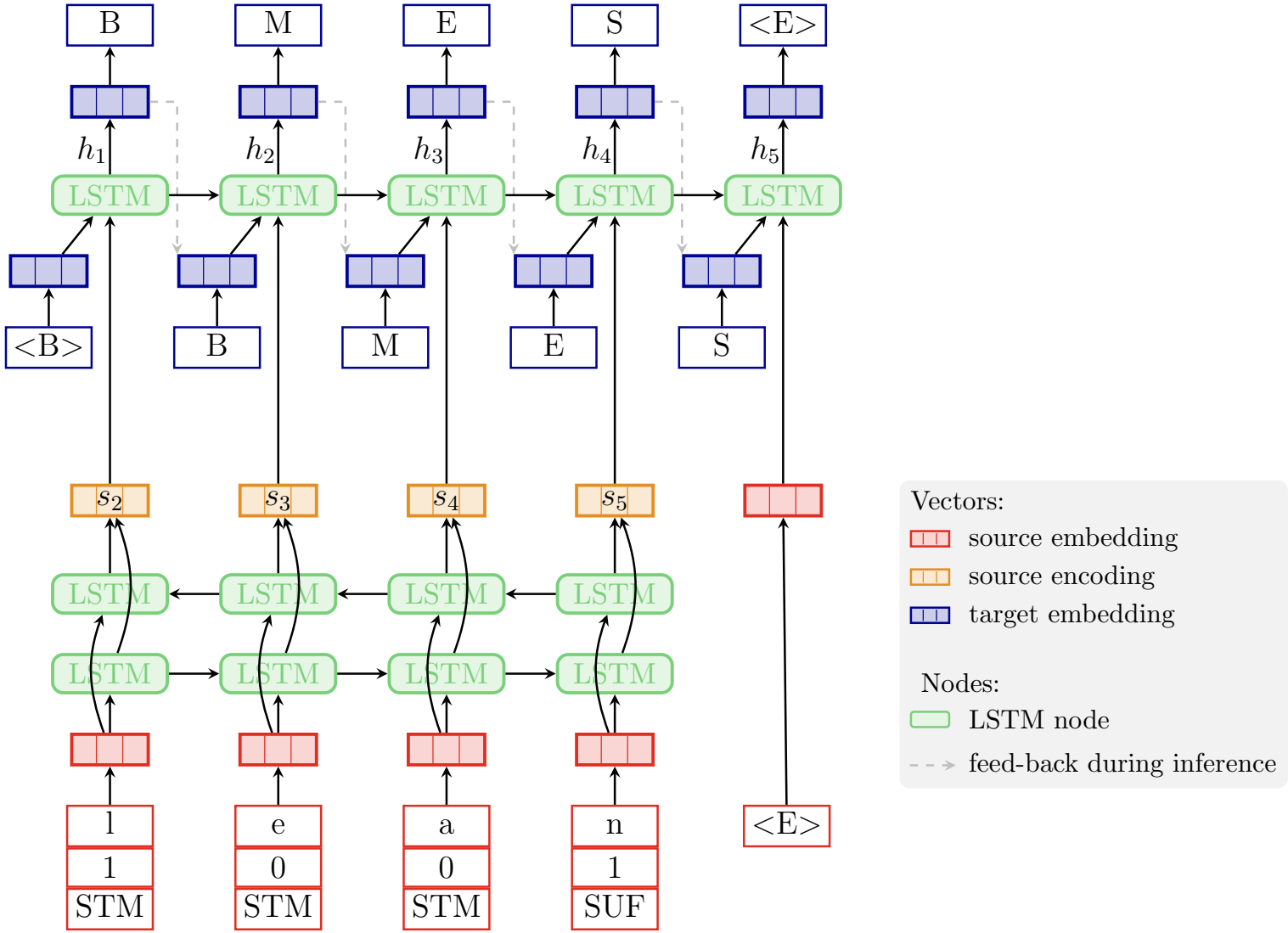
Seq2Seq for morphological segmentation

- Seq2seq model applied for low-resource morphological segmentation (Kann et al., 2018)
- 4 South American polysynthetic languages
  - Mexicanero, Nahuatl, Wixarika and Yorem Nokki
- Training data: 427 to 665 word forms per language
- Multi-task training to address the copying problem

Models: Seq2Seq



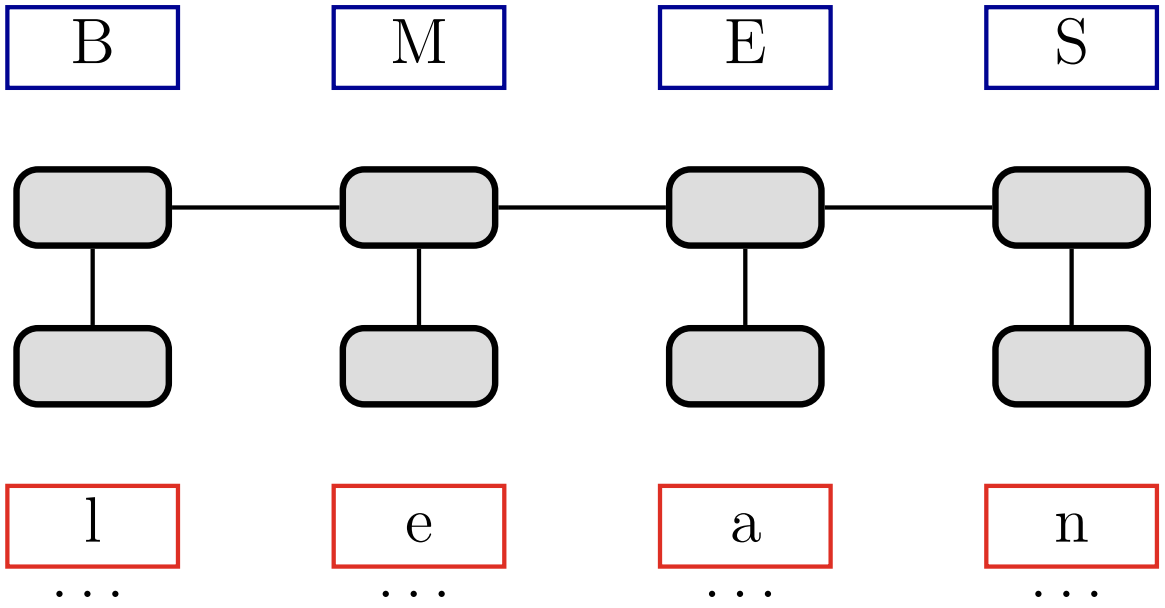
Models: Neural sequence tagger



Models: Neural sequence tagger

- Much smaller output vocabulary.
- Synchronous with the input, no attention needed.
- Smaller optimal network (layers, vector dimensions).
- NST has only 5% of the number of parameters of the seq2seq model.

Models: Conditional random field



Data:

source feature

target label

Nodes:

CRF node

Data and prior state of the art

- Annotations for North Sámi morphological segmentation (Grönroos et al., 2016).
- Collected using an active learning procedure.

Purpose	Subset	Component	Word types	Labels
Training	Unlabeled	FlatCat	691190	No
Training	Feature train Main train	FlatCat System	200 844	Yes Yes
Development		Both	199	Yes
Testing			796	Yes

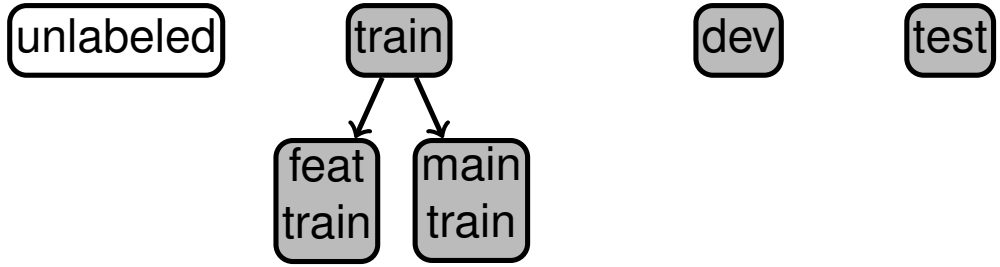
Table: Subdivision of data sets, with size in word types.

Feature set augmentation

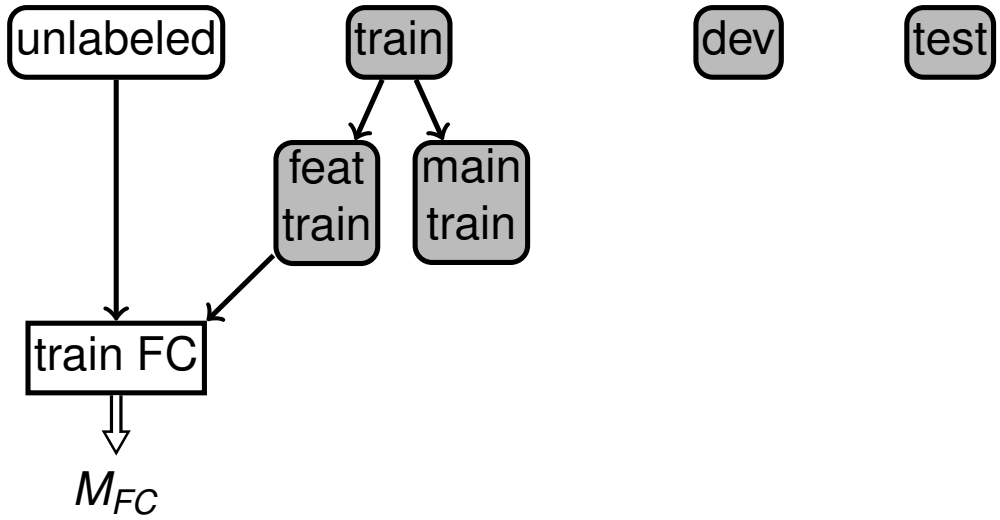
- Feature set augmentation approach
  - (Ruokolainen et al., 2014).
  - Combines the strengths of generative and discriminative models.
- Predictions of generative model (Morfessor FlatCat) used as input features to train a discriminative model.

Training procedure

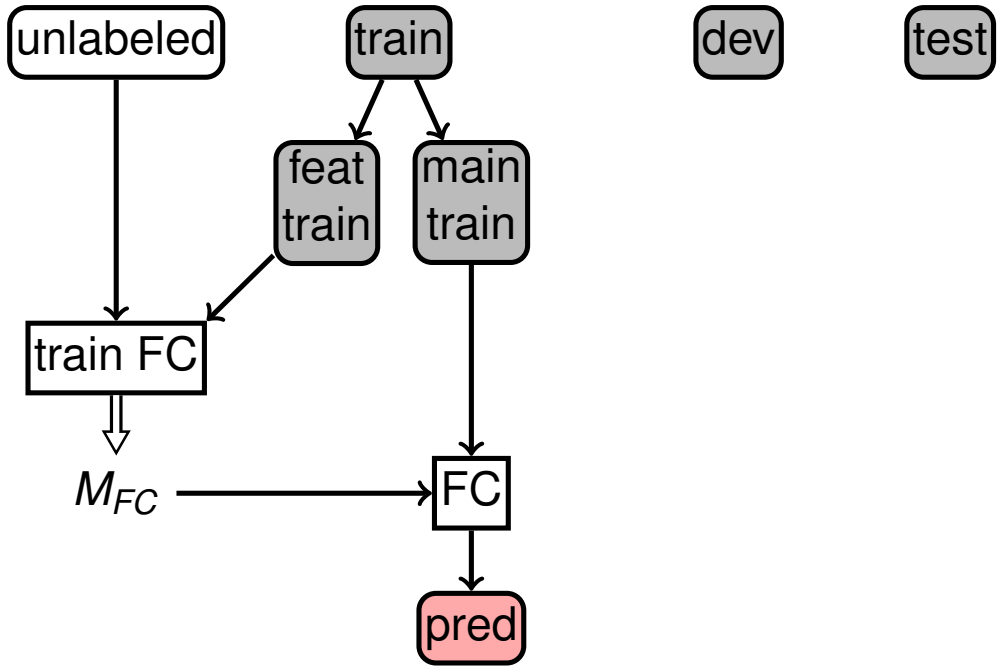
Training procedure



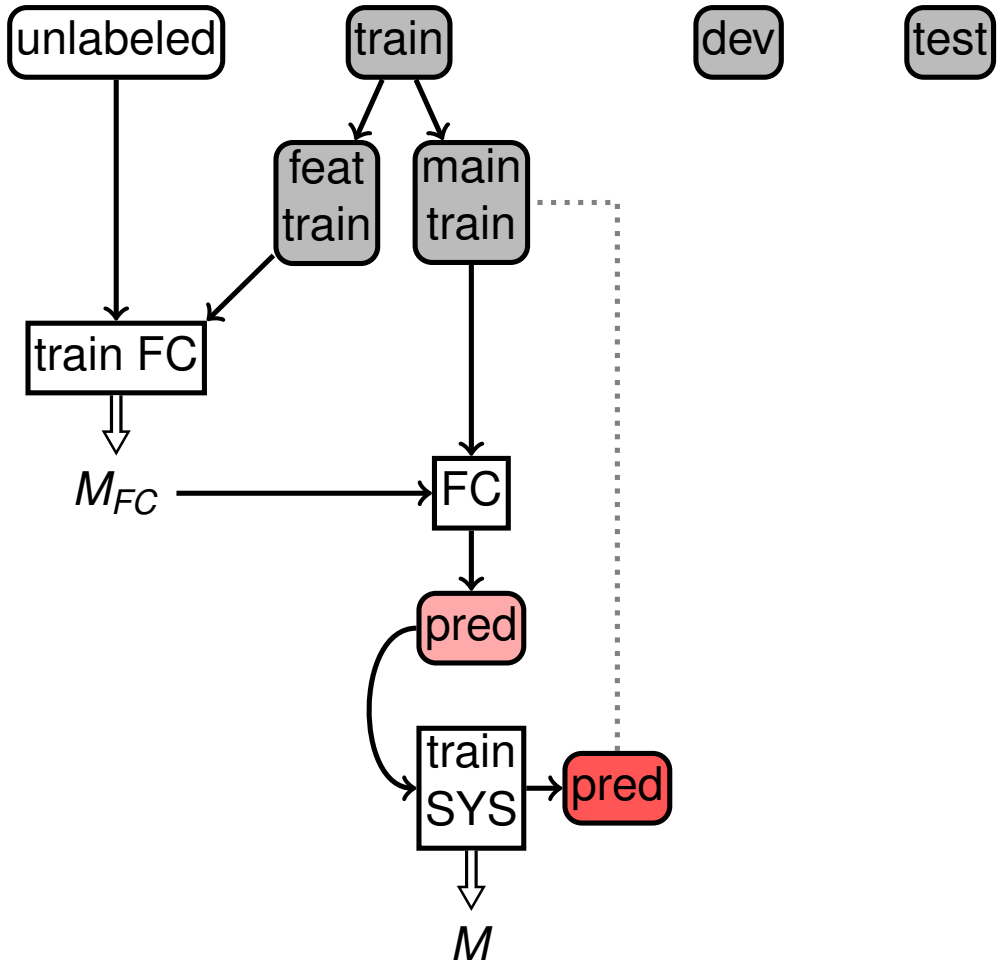
Training procedure



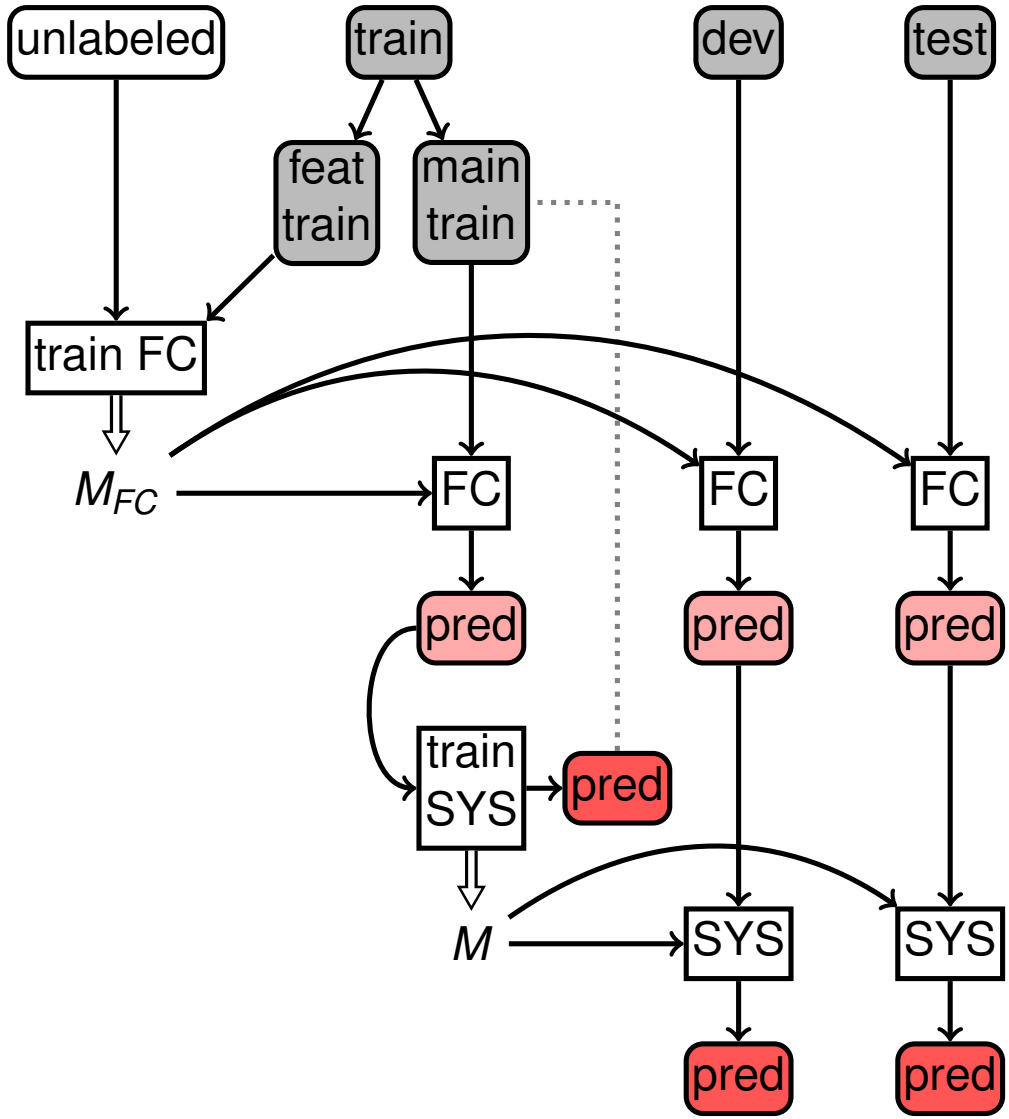
Training procedure



Training procedure



Training procedure



Results

system	Pre	Rec	F_1	w-acc
FlatCat (200 words)	78.20	77.60	77.90	57.20
Seq2seq (s)	86.94	78.62	82.54	64.60
NST (s)	83.26	83.92	83.58	69.12
CRF (s)	**87.70**	83.30	85.40	69.30
FlatCat (full)	74.30	84.10	78.90	61.80
Seq2seq (ss)	87.66	80.16	83.72	68.36
NST (ss)	84.28	**85.58**	84.94	71.02
CRF (ss)	86.30	85.20	**85.70**	**71.10**

Table: Results on the test set. Boundary precision, recall, and F_1-scores, together with word-type level accuracy. Higher is better. Averages of 5 runs.

Results: category patterns

- Seq2seq has the best performance for the STM-pattern.
- CRF is good at modeling the boundaries of suffixes.
- Fully supervised CRF is poor at splitting compound words.
  - Low STM+STM recall.
  - Alleviated by the FlatCat features.
- The neural sequence tagger is good at modeling the ends of stems.

Examples

- correct
  - dollo + juvvo
  - datte
  - goalmmá + t
  - soga + s
  - daga + h + ii
  - orpma
  - dorski
  - moai + t
  - ovdánahttá
  - muital + ii
- incorrect
  - ravga + t (ravg + at)
  - reahpen + is (reahpeni + s)
  - dakkára (dakkár + a)
  - deaivida (deaivid + a)
  - báhtu + i (báhtui)
  - bivttasskábi + i (bivttas + skábi + i)
  - bellodat (belloda + t)
  - johkagátti (johka + gátti)
  - vuvd + ojuvvo (vuvdo + juvvo)
  - eanan + vuou (eana + n + vuou)

Summary

- Semi-supervised sequence labeling is an effective way to train a low-resource morphological segmentation system.
- We improve 8.6% compared to using FlatCat directly.
- Neural methods are applicable to low-resource settings, but didn't outperform CRF.
- All four systems are language-independent and have open-source implementations.
- Recommendation: Use FlatCat + CRF.

Bonus data release

`http://research.spa.aalto.fi/speech/data_release/north_saami_active_learning/systems_agree.tar.gz`

words	file
352606	systems_agree
310224	systems_disagree
300	systems_disagree.300.ifsubstrings_5n

Bibliography

Stig-Arne Grönroos, Katri Hiovain, Peter Smit, Ilona Erika Rauhala, Päivi Kristiina Jokinen, Mikko Kurimo, and Sami Virpioja. Low-resource active learning of morphological segmentation. *Northern European Journal of Language Technology*, 2016.

Katharina Kann, Manuel Mager, Ivan Meza, and Hinrich Schütze. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of NAACL 2018*, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. Painless semi-supervised morphological segmentation using conditional random fields. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 84–89, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, June 2017. URL `http://arxiv.org/abs/1706.03762`. arXiv: 1706.03762.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Q Le, Y Agiomyrgiannakis, R Clark, and R. A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010, August 2017.