Fribourg, 4th Sep 2025

Improving the Quality of Morphological Segmentation using Self-Training Methods

Michal Olbrich
Zdeněk Žabokrtský

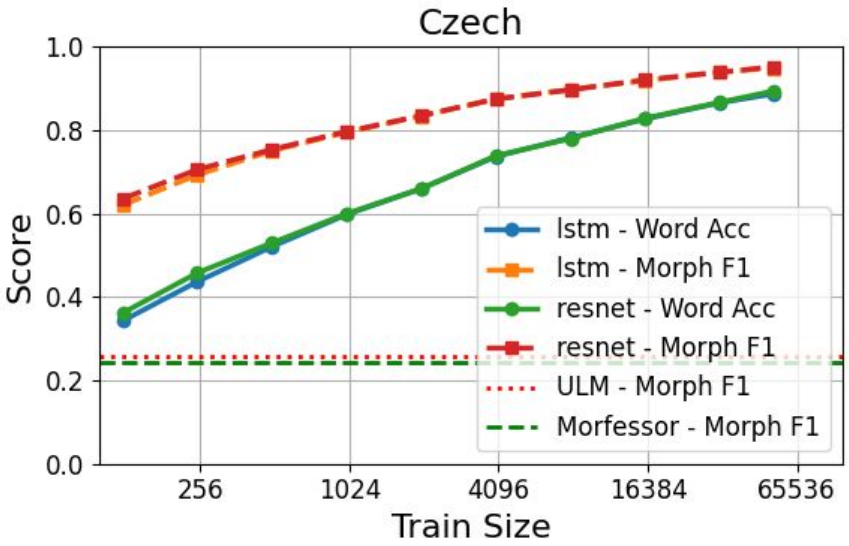
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

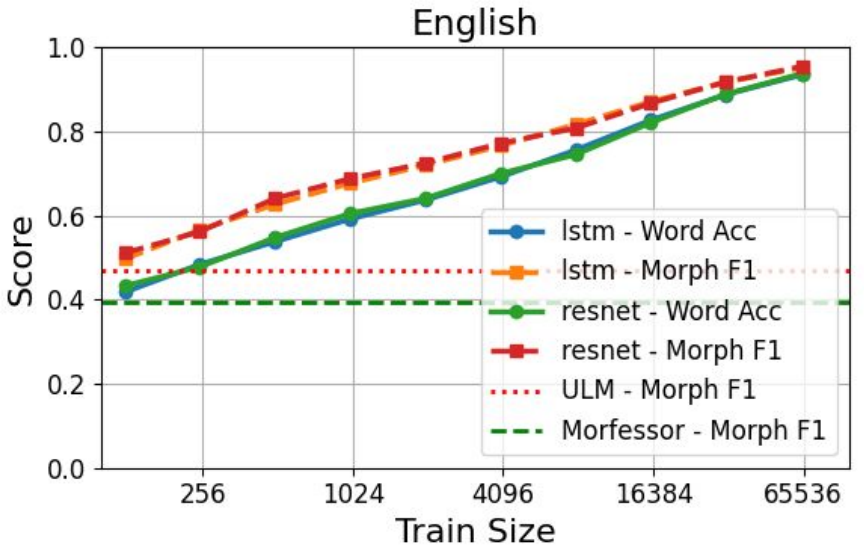
Outline

1. Motivation - inconsistencies in the datasets
    - Goal: improve automatic segmentation quality (Morph F1/Word accuracy), but plateauing scores suggest possible inconsistencies in training/test data
2. Linguistic perspective: ambiguous and irregular word formation
3. Methodology: What is Self-Training
4. Results
    - Detection of inconsistencies
    - Asymmetrical Noise Injection
5. Conclusion

Motivation

- Simply increasing the dataset size is not the answer
  - **Sigmorphon 2022 Shared task on morphological segmentation** had train size for English dataset of 500 kw and the results were worse than for our models trained on much smaller MorphoLex dataset (65 kw) - **93.6%** Morph F1 for the winning system vs. **95.5%** Morph F1 for ours





Morphologically segmented resources

- Large semi-automatically gathered resources
  - UniMorph: mostly inflectional morphology, 182 languages (Batsuren et al.,2022)
  - MorphyNet: primarily derivational morphology, 15 languages (Batsuren et al., 2021)
- Linguistically accurate dictionaries (limited number of language)
  - Russian: Tikhonov (1990)
  - Czech: Slavíčková (1975) - problem with digitization of old “paper” dictionaries
  - Slovak: Ološtiak (2015)
  - English, French, German : (MorphoLex, CELEX)
- Many problems in the process of harmonization of such resources
  - Different formats
    - Surface vs. canonical: *funniest* → *funn-i-est* vs. *fun-y-est*
    - Derivational networks
  - Completeness of segmentation, inner inconsistencies, only selected POS (verbs)...
  - Project **Universal Segmentations** - 32 languages, various sources (Žabokrtský et al., 2022)

Difficulties of morphological segmentation

Original Word	Segmented Form	Gloss
pek**árna**	pek-**ár-n-a**	bakery
tisk**árna**	tisk-**árn-a**	printing house
ocel**árna**	ocel-**árn-a**	steel mill
kas**árna**	kas**árn-a**	barracks
kav**árna**	kav-**árn-a**	coffee shop
tov**árna**	tov**ár-n-a**	forge

- **“-árna”** marks places of crafts/production, but segmentation varies depending on whether it derives from the agent (person) or the activity

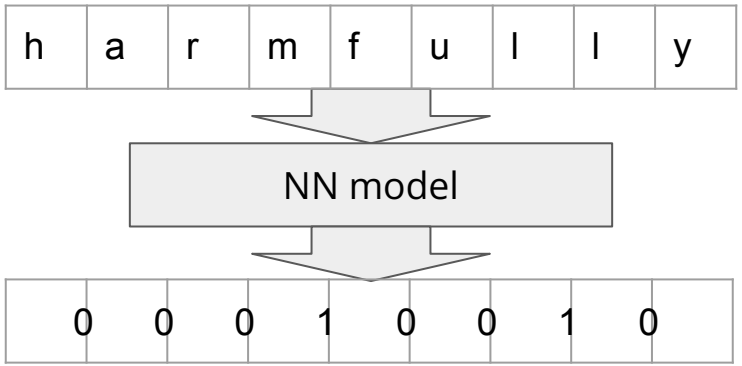
Datasets

- Czech dataset - SIGMORPHON 2022 Shared Task on Morpheme Segmentation dataset expanded with some additional other source
- Slovak - Retrograde Morphemic Dictionary of Slovak (Ološtiak)
- English - Universal Segmentations converted dataset from MorphoLex

Language	Train Size	Test Size	Avg. Boundaries/Word	Avg. Word Length
Czech	52,458	4,000	2.6	8.1
English	64,623	4,000	1.2	8.3
Slovak	65,430	4,000	2.9	8.6

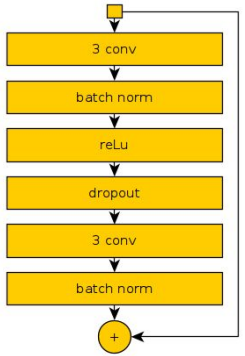
Neural network models

- Sequence to sequence models
- Characters on the input
- Segmentation boundaries on the output



Neural network models

- Convolutional vs. recurrent
  - Both perform very similarly, but recurrent tend to over fit more, while convolutional seems to better generalize
  - Convolutional
    - 15 layers of 1D ResNet blocks (convolutional blocks with skip connections), with kernel size 3, number of filters 240
  - Recurrent
    - 1D ResNet blocks on the input followed by 1 biLSTM block of width 600
  - 35 epochs, Adam with cosine decay, 5 warm up epochs, initial learning rate: 0.05, dropout: 0.1, label smoothing 0.05, Binary cross-entropy loss, batch size 2 for 125 words, doubling with the train size



- Comparison to unsupervised methods
  - ULM
  - Morfessor

Experiment - detection of inconsistencies

- Some of the boundaries are misplaced
  - Annotation bias: Missing boundaries are more common than extra ones
    - Task design: annotators only mark boundaries, but don’t confirm non-boundaries
    - Human tendency: annotators are cautious → better to skip a boundary than risk a wrong one
    - Ambiguity: some segmentations are genuinely debatable → leads to under-marking
- Manual revision of large datasets is time demanding, some errors might be missed
- Prediction on the train set can reveal annotation inconsistencies as well as linguistic irregularities.

Self Training

- Use a model's own predictions on unlabeled data as additional training data
- **Pseudo-labels**: The model assigns "temporary" annotations to unlabeled morphological boundaries
- **Manifold Assumption**: Words that are close (similar in form/structure) should share similar morphological segmentations
- **Task-specific issue**: In sequence labeling, 0 may mean true "no boundary" or simply missing/unlabeled data
- **Self-Improvement Loop**: Train → predict → add confident predictions → retrain
- **Risk**: Model can reinforce its own mistakes if pseudo-labels are wrong
- **Evaluation challenge**: Without true gold annotations, it's hard to measure improvement reliably

Detection of inconsistencies - Czech

- Main language of interest - multiple sources - different annotators

#	Word	Gold Segmentation	Predicted Segmentation	Gloss
1	divadlo	div-a-dlo	div-a-dl-o	theater
2	vojenského	vojen-sk-ého	voj-en-sk-ého	military (gen. sg.)
3	polotovar	polo-tovar	pol-o-tovar	semi-finished product
4	prosmyknout	pro-smyk-nou-t	pro-s-myk-nou-t	to slip through
5	vzájemné	vzájem-né	v-zá-jem-né	mutual (nom. pl. n.)
6	půlmiliardový	půl-miliard-ov-ý	půl-mili-ard-ový	worth half a billion
7	záplata	záplat-a	zá-plat-a	patch
8	akcionář	akcion-ář	akci-on-ář	shareholder

Detection of inconsistencies - English

- English MorphoLex dataset is actually already really good
    - Word Accuracy of 94% - meaning every 20th word was wrongly predicted
    - Not much space for improvement - mostly under-segmented Latin/French borrowings

#	Word	Gold Segmentation	Predicted Segmentation
1	incurability	incur-abil-ity	in-cur-abil-ity
2	extravagantly	extravagant-ly	extravag-ant-ly
3	disconsolate	disconsolate	dis-consolate
4	atomically	atom-ic-ally	atom-ic-al-ly
5	unless	unless	un-less

Detection of inconsistencies - Slovak

- Professionally created dictionary, nevertheless the model was still able to detect inconsistencies
- Similarly to English dataset - little room for improvement - 95 % Word Accuracy (98% Morph F1)

#	Word	Gold Segmentation	Predicted Segmentation	Gloss
1	rozhodne	rozhod-ne	roz-hod-ne	decides
2	frajerkárstvo	frajer-k-ár-stv-o	fraj-er-k-ár-stv-o	philandering
3	vyparatiť	vyparat-i-ť	vy-parat-i-ť	to make mischief
4	prosperovať	prosper-ov-a-ť	pro-sper-ov-a-ť	to prosper
5	dvíhačka	dvíh-a-čk-a	dvíh-a-č-k-a	jack (lifting device)

Detection of Incorrectly Annotated Segmentations

- In the first iteration of this experiment, the model produced different segmentations for 1,168 words compared to manual annotations, resulting in a train Word Accuracy of 97.9%. Among these predictions, annotators corrected 328 words (27.5%) from which 278 were exactly predicted by the model. Test Word Accuracy was **87.3%.**
- The total number of added boundaries by annotators was 433, from which 377 were detected by the model and on top of that **56 were added** by the annotators.
- In contrast to that, **only 6 segmentation boundaries were removed** by the annotators.
- After correcting those 439 erroneous segmentation borders and updating the data set while maintaining the same train-test split, the Word Accuracy on the test set improved to **88.1%**, marking an increase of **0.8%**.
- Repeating the same experiment on the corrected dataset led to a further improvement of word accuracy to **88.8%**.

Recovery of Missing Boundaries under Noisy Supervision

- To simulate the model's ability to correct erroneous or inconsistent annotations, we conducted controlled experiments with **asymmetric label noise injection**, where 5% or 10% of segmentation boundaries were randomly removed from the training data

Recovery of Missing Boundaries under Noisy Supervision

$$\text{BRA} = \frac{\#\text{Recovered boundaries}}{\#\text{Removed boundaries}}$$

5 % removed

Language	15 Epochs		25 Epochs		35 Epochs	
	BRA	Precision	BRA	Precision	BRA	Precision
Czech	0.91	0.81	0.83	0.91	0.59	0.96
Slovak	0.93	0.89	0.86	0.96	0.61	0.99
English	0.92	0.64	0.76	0.88	0.44	0.94

10 % removed

Language	15 Epochs		25 Epochs		35 Epochs	
	BRA	Precision	BRA	Precision	BRA	Precision
Czech	0.90	0.90	0.87	0.92	0.77	0.96
Slovak	0.94	0.94	0.91	0.97	0.83	0.99
English	0.87	0.86	0.75	0.93	0.66	0.94

Conclusion

- Self learning / pseudo labeling could be used to improve annotation consistency or to recover missing boundaries.
- Particularly useful for detecting inconsistencies in large datasets, where manual revision is time-consuming.
- Further research is needed to explore optimal parameters for fully automated unsupervised Self-Improvement Loop