Does Topological Ordering of Morphological Segments Reduce Morphological Modeling Complexity?

A Preliminary Study on 13 Languages

Andreas Shcherbakov
The University of Melbourne
scherbakov.andreas@unimelb.edu.au

Kat Vylomova
The University of Melbourne
vylomovae@unimelb.edu.au

Overview

- We propose augmenting the inflection model with segmentation.

We suggest that

- annotated morphological segmentation can significantly improve the generalization ability.
- such task is easier to solve than the reinflection task in its classical setting, especially in agglutinative languages.
- the reinflection task can be formalized as a classification task rather than a string-to-string transduction task.
  - reduction of the search space
  - enhancing the model's robustness to data sparsity

Dataset

Morphynet: a large multilingual database of derivational and inflectional morphology.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Languages:

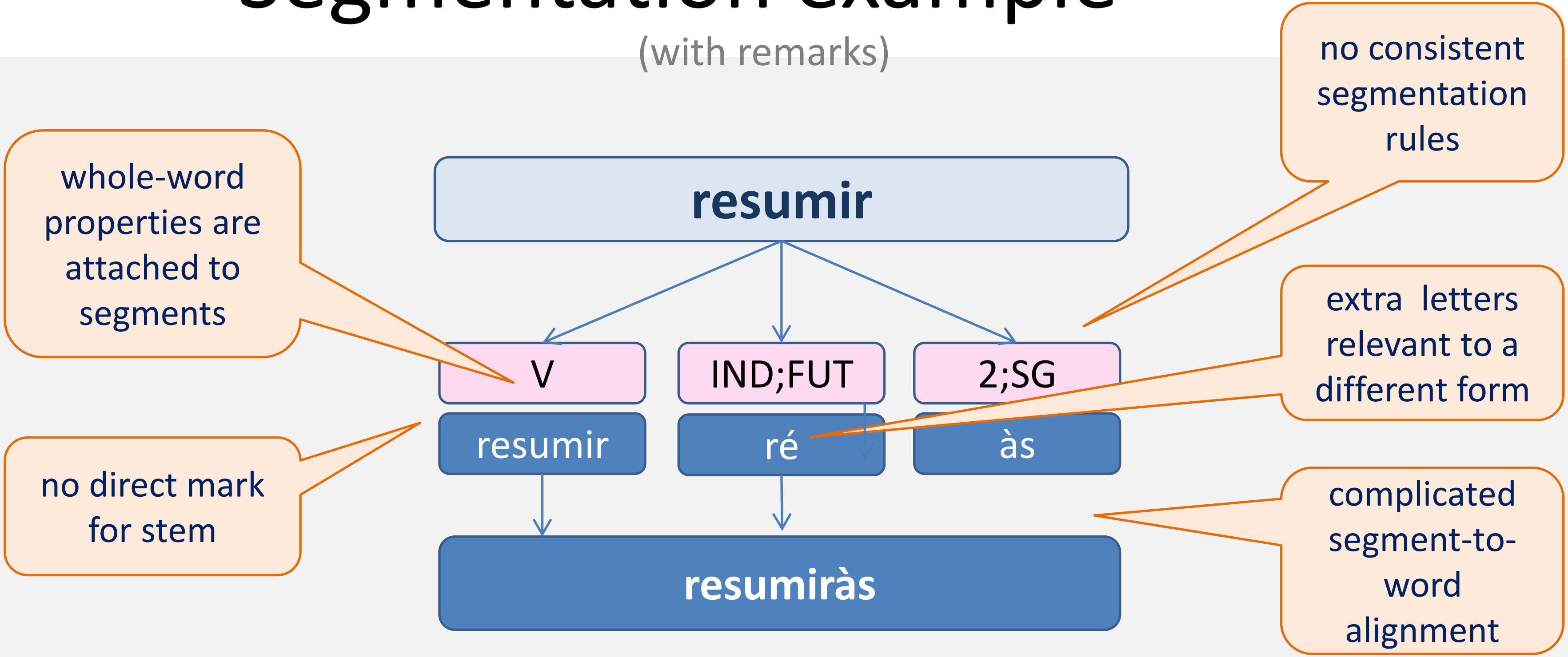
- Catalan (cat)
- Czech (ces)
- German (deu)
- English (eng)

- Finnish (fin)
- French (fra),
- Hungarian (hun),
- Italian (ita)

Mongolian (mon)
Portuguese (por)
Russian (rus)
Spanish (spa)
Swedish (swe)

Segmentation example

(with remarks)



whole-word properties are attached to segments

no direct mark for stem

no consistent segmentation rules

extra letters relevant to a different form

complicated segment-to-word alignment

Global segment order

- We found that segment-wise tagsets are strictly ordered globally within a given language: mapping $t \longrightarrow p$ is possible, where
  - t is distinct segment-wise tagset
  - p is distinct number
  - segment with lower p comes earlier in a word
- Therefore, no seq-to-seq is needed to predict segments – a plain classifier should do the job.
  - ➢ Solution space reduction, smaller training sets.

Segments to word

- Seq2seq is still needed to combine segments into words

➢ Fortunately, it’s a rather easy task:

❖ A hard attention model (Aharoni and Goldberg, 2017) predicts segments “gluing” into a word nearly perfectly (see the table)

- German was the only exception due to compounding.
- No tags were provided to the model

Examples of transformations

- -ar → -u- in Spanish:

  catalogar → V|IND;PRS;1;PL

  catalog**ar|em** → catalog**uem**

- removal of adjacent duplicate letters
- replacement of certain adjacent letter combinations at segment boundaries

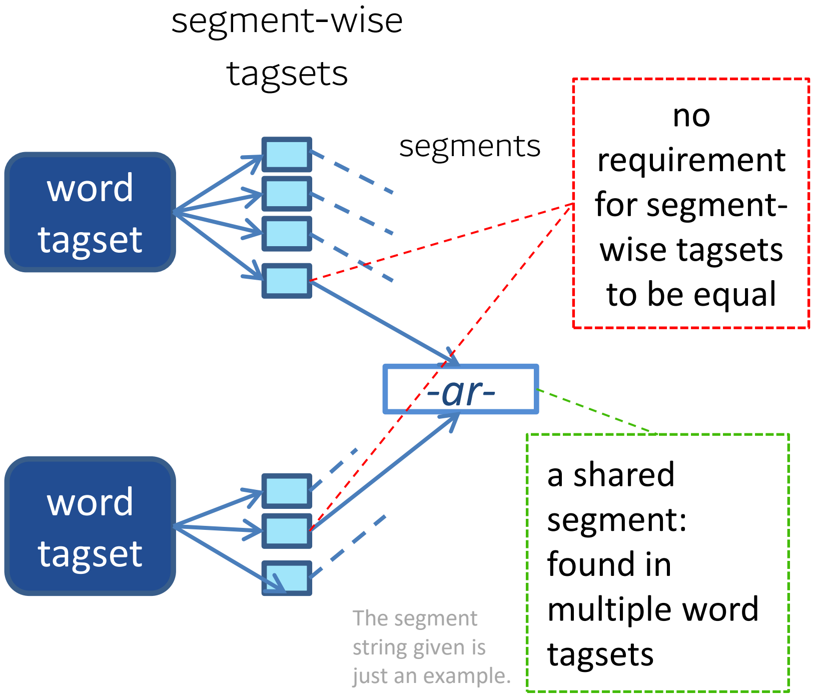
  (Czech) čtverec → N;SG|INST;MASC;INAN

  čtverec|em → čtvercem

cat	ces	deu	eng	fra	hun	Ita	mon	por	swe
.99	.98	.89	.99	.99	.98	.99	1.00	1.00	.98

Challenge of agglutinativity

- Agglutinative languages are the most hard for the re-inflection task due to a model's lack of generalization.
- Q: At which probability (*recall*) a correct affix segment list for an unobserved word tagset can be reconstructed as combination of segments observed for other tagsets?


segment-wise tagsets
segments
word tagset
word tagset
-ar-
no requirement for segment-wise tagsets to be equal
a shared segment: found in multiple word tagsets
The segment string given is just an example.

Tagset composability

- *Composability* = percentage of “composable” word tagsets.
    - “composable tagset” = one that shares all representing segments with some other tagsets.

(We ignore tagsets which include tags not seen in any other tagset).

cat	ces	deu	eng	fin	fra	hun	ita	por	rus	spa	swe
.85	1.00	.96	.50	1.00	.52	.88	.55	.55	.98	.96	.97

High composability values for agglutinative languages suggest utility of segmentation for prediction of morpheme combinations.

(They may be high for some fusional languages as well)

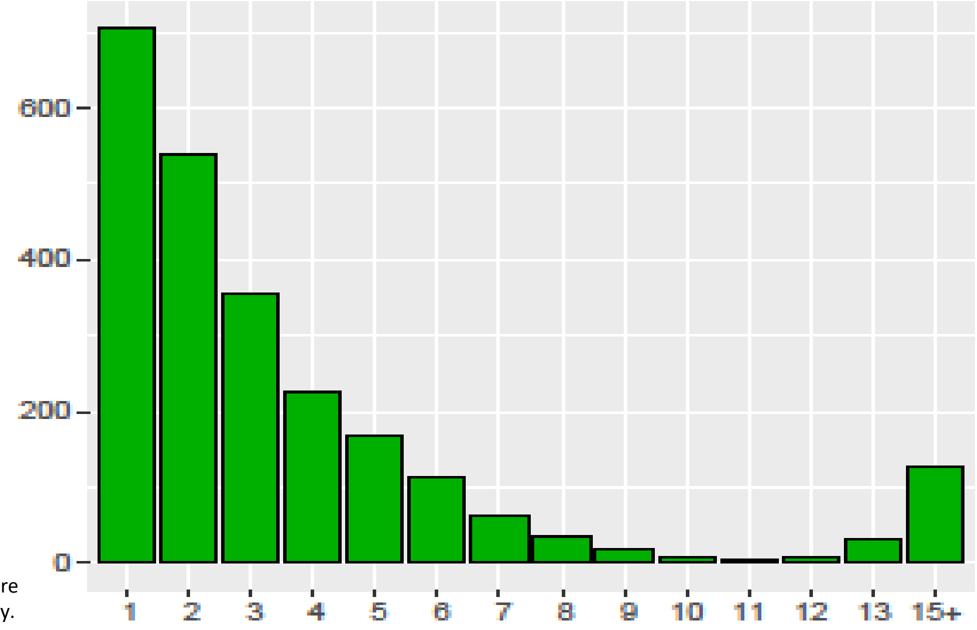
Decomposing word tagsets

- *Word tagset -> segment-wise tagsets -> segment strings* inference is too noisy in MorphyNet.
- Still, luckily, a more direct one works pretty well: word tagset -> segment strings.

➢ We recommend the latter way despite it’s “less natural”.

➢ We observed a low entropy distribution of distinct segment combination per word tagset.

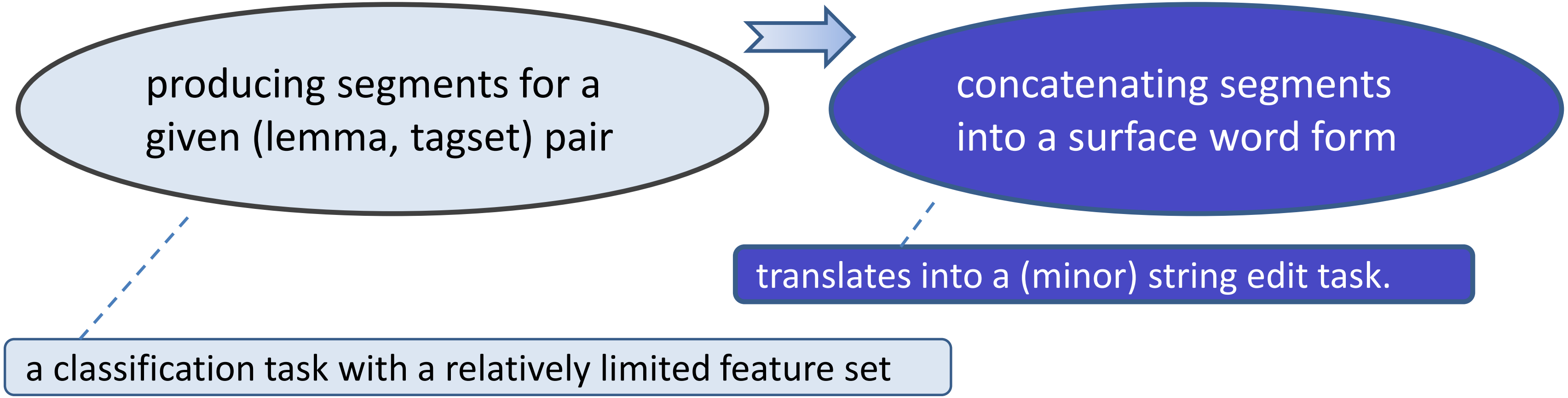
➢ Meanly, only a few options of target segment combinations (usually less then 4) per word tagset were observed in the dataset.



A frequency distribution for the number of different morphological segments per tagset. Here we consider distinct (language, tagset) pairs.Affix (non-stem) segments were considered only.

Suggestions from the experiments

The usage of morphological segmentation dataset enables principal reduction of the complexity of the morphological inflection task by breaking it into:



producing segments for a given (lemma, tagset) pair

concatenating segments into a surface word form

translates into a (minor) string edit task.

a classification task with a relatively limited feature set

! Segmentation resources are only available in few languages

! Segmentation conventions are ambiguous and need standardization

Prospective tasks

- the ability to generalize to unseen grammatical tag combinations (Kodner et al., 2022)
- to better account for phonotactics
- application to smaller training sets for under-resourced languages
- finding balance between latent and explicit segmentation

Conclusions

- We conducted a series of experiments with morphological segmentation and demonstrated that annotated segment sequences may significantly **simplify** the prediction of inflected forms.
- We outlined that inflection task can be transformed from sequence-to-sequence into a classification task, with better capacities to address language agglutinativity challenges.

Thank you!