Morphological Segmentation Can Improve Syllabification

Garrett Nicolai, Lei Yao, and Greg Kondrak

Department of Computing Science

University of Alberta

How can morphological information help syllabification?

- Syllabification is the process of breaking a word into syllables.
- Typically, a phonemic process; when splitting orthographic representations, syllabification is called *hyphenation*.
- Useful for determining line breaks in documents.
- Can aid pronunciation: proph-et, up-hold.
- Many syllable breaks coincide with morphological breaks: black-board, re-fut-a-ble, hold-ing.
- Highly accurate systems, such as Bartlett et al.(2008) make mistakes that could benefit from morphological information: *hol-dov-er, *coad-ju-tors.
- We show that morphological information can help syllabification.
- Somewhat surprisingly, unsupervised methods are as good or better than supervised ones.

Morphological Features

- Syllabification System of Bartlett et al. (2008) serves as a baseline.
- Structured SVM predicts tags in sequence.
- A tag is predicted for each letter in a word.
- Tags follow Numbered NB format.
- Features include *n*-grams around focus characters, from unigrams to 5-grams.
- Orthographic features are supplemented with morphological annotation.

Morphological Annotation

Word:	syllabify
Syllables:	syl-lab-i-fy
Morphs:	syllab+ify

Letters:	s	y	l	l	a	b	i	f	y
Tags:	N1	N2	B	N1	N2	B	B	N1	N2

Morphological features

s y l l a b i f y →

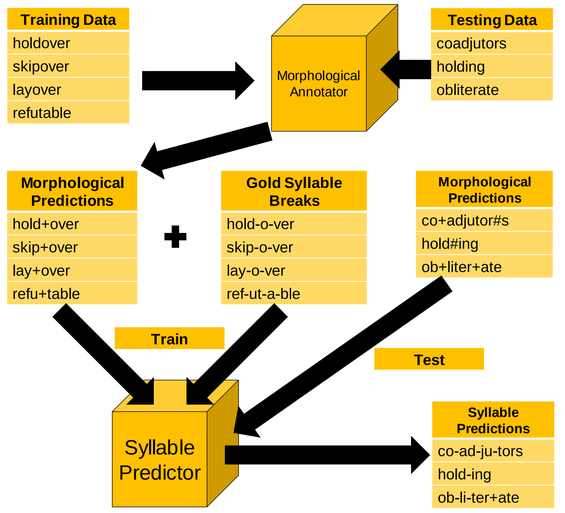
Unigrams	Bigrams	Trigrams	4-grams	5-grams
s.._ y._ l_ l_ _a _*b _**i _***f _****y	^s**_ sy._ yl_ l̪l̪ l̪a _ab _*b+ _**+i _***if	^sy*_ syl_ yl̪l̪ l̪la l̪ab _ab+ _*b+i _**+if	^syl_ syl̪l̪ yl̪la l̪lab l̪ab+ _ab+i _*b+if	^syl̪l̪ syl̪la yl̪lab l̪lab+ l̪ab+i _ab+if

Supplementing with Morphological Annotation

Sources of Morphological Annotation

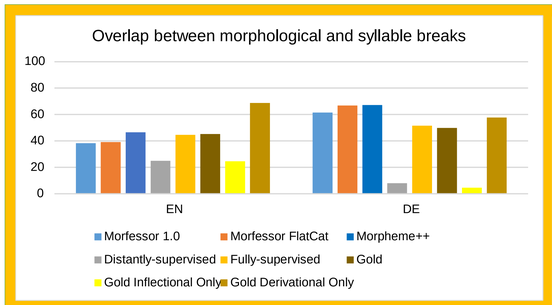
- Hand Annotated Lexicons
  - Very accurate..
  - Require Expert Knowledge.
  - Rare and Expensive to create.
  - What to do about unannotated forms?
- Fully-supervised Systems
  - Still require some annotated data.
  - Can predict unseen forms.
- Semi- and distantly-supervised systems
  - Less annotation required.
  - Generally less accurate than fully-supervised systems.
- Unsupervised systems
  - No annotation required.
  - Not as accurate as supervised systems.

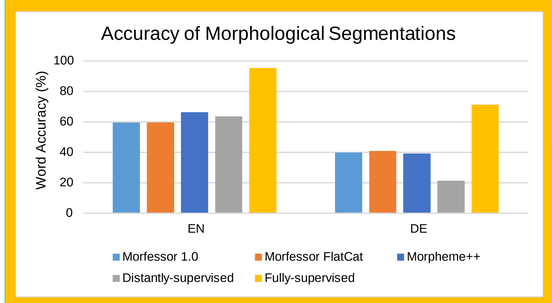
Annotation Process

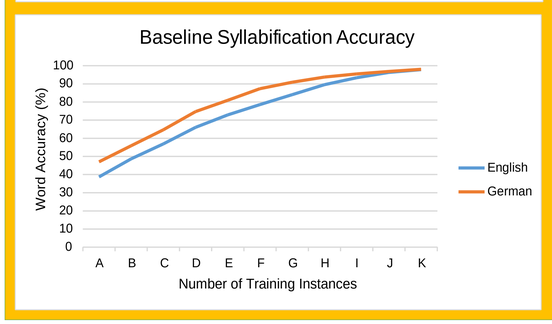


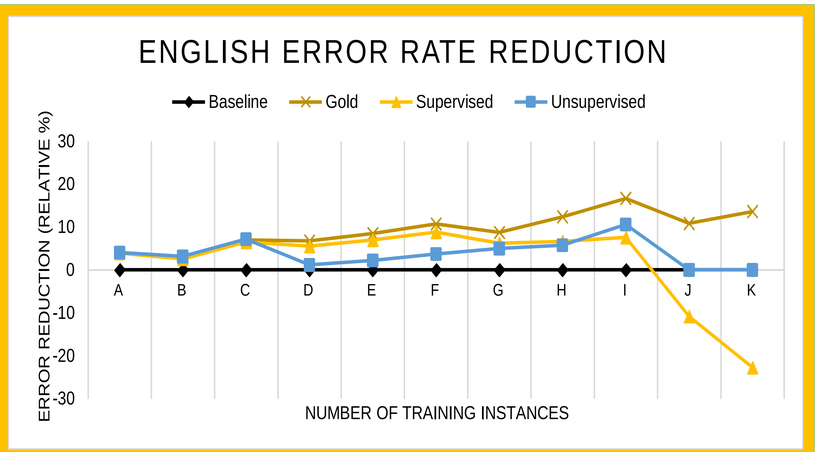
Experimental Setup

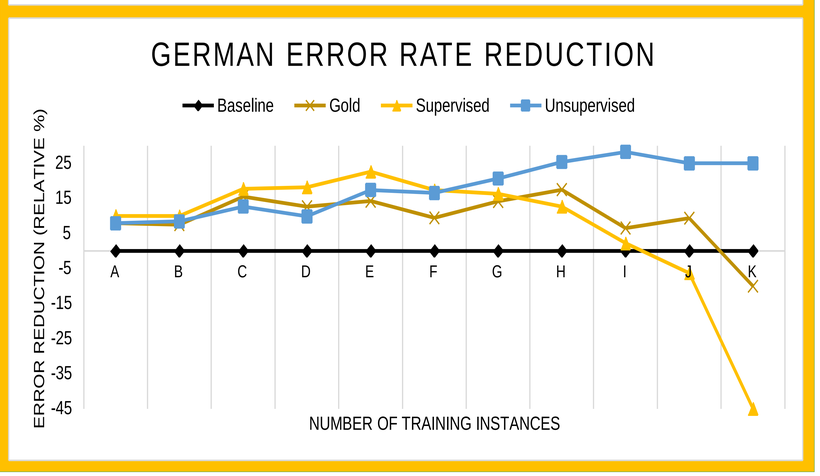
- We consider two languages: English and German.
- Gold syllable breaks are taken from CELEX.
- Our baseline is supplemented with morpheme breaks predicted by several methods:
  - Gold breaks from CELEX:
    - inflectional breaks only.
    - derivational breaks only.
    - all breaks (Both)
  - Breaks predicted by a fully supervised system.
  - Breaks predicted by a distantly-supervised system.
  - Three unsupervised systems:
    - Morfessor 1.0
    - Morpheme++
    - Morfessor FlatCat











Discussion

- The fully-supervised method provides a benefit at smaller training sizes, but harms accuracy at larger training sizes.
- The unsupervised methods continue to improve accuracy as training size increases.
- The supervised system misses some compounds: since compound breaks are almost always syllable breaks, this hurts the supervised system's performance..
- The unsupervised methods are able to identify segments that are not productive affixes, such as "ob" in *obliterate*; these segments are often syllables of their own.

Conclusions

- Morphological information can aid syllabification.
- Unsupervised methods often out-perform supervised ones, and can rival gold annotation.