

Non-parametric Bayesian models for computational morphology

Dissertation defence

Kairit Sirts

Institute of Informatics
Tallinn University of Technology

18.06.2015

Outline

1. NLP and computational morphology
2. Why non-parametric Bayesian modeling?
3. Thesis Claims
4. Model 1: joint POS tagging and morphological segmentation
5. Model 2: weakly-supervised morphological segmentation
6. Model 3: morphosyntactic clustering using distributional and morphological cues
7. Future work

Natural language processing

- Human-human interaction



- 1-2 languages



World's languages

Big languages

Mandarin
English
Spanish
Hindi/Urdu
...

1750

Less than 1000 speakers

7000
languages

~275

More than 1M speakers

Level of computer support



Language complexity

- Related to morphological complexity

English Nouns

4 inflected forms

	Singular	Plural
Nom	bird	birds
Gen	bird's	birds'

Estonian Nouns

28 inflected forms

	Singular	Plural
Nom	lind	linnud
Gen	linnu	lindude
Part	lindu	linde
III	lindu	lindudesse
	...	...

Morphology

- Studies the words' internal structure

Definition 1 (Haspelmath and Sims, pp. 3):

Morphology is the study of the combination of morphemes to yield words.

Morphemes are the smallest meaningful constituents of words.

- disconnections → dis_connect_ion_s

Definition 2 (Haspelmath and Sims, pp. 2):

Morphology is the study of systematic covariation in the form and meaning of words.

- Mutter → Mütter

Computational morphology

- Useful for:
 - machine translation, speech recognition, information retrieval, natural language generation

SPARSITY

- Infrequent words (Zipf's law)
- Fixed size vocabularies

Recognize a word:

disconnection → out of vocabulary
dis, connection → in the vocabulary
disconnection = dis + connection

Computational morphology tasks

- Morphological segmentation
 - Splitting words into morphemes
 - disconnections → dis_connect_ion_s
- Part-of-speech tagging (clustering)
 - Clustering words based on their syntactic function
 - noun, verb, adjective, pronoun, ...
- Morphological analysis
 - Assigning each word a set of morphosyntactic features
 - hallides → hall+des //_A_ pos pl in //

Why non-parametric Bayesian modeling?

- Supervised vs **unsupervised**
 - Enables working with languages lacking annotated linguistic data
- Algorithmic vs **model-based**
 - Probabilistic modeling framework
 - Provides semantics to the model
- Frequentist vs **Bayesian**
 - Frequentist: $P(\text{Data} | \text{Model})$
 - Bayesian: $P(\text{Data} | \text{Model}) * P(\text{Model})$
 - **Non-parametric** priors generate Zipfian distributions

