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



• Human-computer interaction



• 1-2 languages



• 90 languages





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): <u>Morphology</u> is the study of the combination of morphemes to yield words. <u>Morphemes</u> are the smallest meaningful constituents of words.

• disconnections \rightarrow 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 \rightarrow 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 \rightarrow out of vocabulary dis, connection \rightarrow in the vocabulary disconnection = dis + connection

Computational morphology tasks

- Morphological segmentation
 - Splitting words into morphemes
 - disconnections \rightarrow 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 \rightarrow 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(Data|Model)
 - Bayesian: P(Data | Model) * P(Model)
 - Non-parametric priors generate Zipfian distributions

Claim A

For **unsupervised** or **weakly-supervised** learning of natural language structures, it is vital not only to model the known properties of those structures, but also some **regularities** or **patterns** that are **latent**, even if they have no specific meaning in linguistic terms.

Claim B

Unsupervised learning can be improved by integrating different aspects of the same process into the **joint model**; this helps to resolve ambiguities, leading to overall better results.

Joint POS induction and morphological segmentation

Model 1

Joint POS induction and morphological segmentation



Results

- Competitive results in POS induction, tested on 15 languages
- Mediocre results in morphological segmentation, tested on 4 languages
- Assess the joint learning with semi-supervised experiments (Estonian)

	Tags	Segments
Unsupervised	47.6	51.9
Semi-supervised	40.5	44.5

Contributions

- State-of-the-art results in unsupervised POS induction over several languages
- Empirical evidence that morphological information and POS assignments influence each other in the joint learning setting (Claim B).

Weakly-supervised morphological segmentation

Model 2

Weakly-supervised morphological segmentation

- Adaptor Grammars framework (Johnson et al., 2007)
 - Combines probabilistic context-free grammars and non-parametric Bayesian modeling
- Two weakly-supervised methods:
 - AG Select uses model selection
 - Semi-supervised AG
- Comparing morphology grammars:
 - word is a sequence of morphemes
 - with morpheme sub- or super-structures

Grammars for learning morphology

Word \rightarrow Morph⁺

Word \rightarrow Morph⁺ Morph \rightarrow SubMorph⁺

Word \rightarrow Compound⁺ Compound \rightarrow Prefix* Stem Suffix* Prefix, Stem, Suffix \rightarrow SubMorph⁺

Results

- Average F1-scores over four languages (Eng, Est, Fin, Tur)
- Weakly-supervised models use 1000 annotated word types

	Unsupervised	Weakly-supervised
AG MorphSeq	58.0	63.4
AG SubMorphs	63.3	66.1
AG Compounding	62.4	69.8 ¹
AG Select		70.8

¹Turkish excluded

Contributions

- State-of-the-art results in both unsupervised and weaklysupervised morphological segmentation across several languages
- Empirical evidence that grammars modeling additional latent sub- or superstructures perform consistently better than the grammars modeling flat morpheme sequences only (Claims A and B).

Morphosyntactic clustering using distributional and morphological cues

Morphosyntactic clustering using distributional and morphological cues

- Unsupervised clustering model
- Distributional information via word embeddings
- Non-parametric prior using **suffix similarity function**
- Clustering and similarity function learned jointly

Word embeddings

- Trained with neural networks
- clustered as multivariate Gaussian random variables



Results on English

- Good results on English
- Not too impressive results on other languages

Model	# Clusters	Accuracy
K-means baseline	104	16.1
IGMM baseline	55.6	41.0
Our model	47.2	64.0

Contributions

- Empirical evidence that the joint model using both sources of information learns better clusters then the one using distributional information only (Claim B)
- Showing that the non-parametric model allowed to choose the number of morphosyntactic clusters freely makes a reasonable choice in English (Claim A)

Future research

- Study the relations between suffixes and (morpho)syntactic categories in morphologically complex languages
 - Current models probably biased to English
- Combine the models together
 - Use Adaptor Grammar segmentation in the joint POS induction and segmentation model
 - Combine the two syntactic clustering models
 - Use learned suffixes as features in the morphosyntactic clustering model
- Apply the segmentation models to more languages

Conclusions

- Three models of computational morphology
- Defined in non-parametric Bayesian framework
- Unsupervised or weakly-supervised
- All employ joint learning in different ways and demonstrate that it is beneficial
- Demonstrate the utility of modeling additional latent structures

Contributions

- Joint POS induction and morphological segmentation
 - State-of-the-art results in unsupervised POS induction over several languages
 - Morphological information and POS assignments influence each other in the joint learning setting (Claim B).
- Weakly-supervised morphological segmentation
 - State-of-the-art results morphological segmentation across several languages
 - Modeling latent sub- or superstructures are helpful for learning morphological segmentations (Claims A and B).
- Morphsyntactic clustering using distributional and morphological cues
 - The model using both sources of information learns better clusters then the one using distributional information only (Claim B)
 - Showing that the non-parametric model allowed to choose the number of morphosyntactic clusters freely makes a reasonable choice in English (Claim A)

Question 1

A question regarding the joint POS induction and morphological segmentation model: One innovation of your model over prior work is the ability to automatically learn the number of tags by using the infinite HMM. What do you think would the impact to your model's performance be if you used a fixed finite number of tags instead, using Dirichlet priors?

Question 2

Regarding the model for morphological segmentation using Adaptor Grammars: In the Adaptor Grammars framework, it was difficult to introduce a weighting factor for a small set of labeled data by simply including each labeled word in the dataset multiple times. What do you think about using weights for the labeled words when computing the posterior grammar, after training? Would that achieve the goal of giving higher weight to observed segmentations?

Adaptor Grammars

Word \rightarrow Morphs Morphs \rightarrow Morph Morphs Morphs \rightarrow Morph <u>Morph</u> \rightarrow Chars Chars \rightarrow Char Chars Chars \rightarrow Char \rightarrow s Char \rightarrow i Char

. . .

PCFG: $P(Word \rightarrow sing_ing) = P(Word \rightarrow Morphs)$ $\times P(Morphs \rightarrow Morph Morphs)$ $\times P(Morph \rightarrow Chars)$ $\times P(Chars \rightarrow Char Chars)$ $\times P(Char \rightarrow s) \dots$

Adaptor Grammar:

 $P(Word \rightarrow sing_ing) = P(Word \rightarrow Morphs)$ × P(Morphs \rightarrow Morph Morphs) × P(Morph \rightarrow sing) × P(Morph \rightarrow ing)

Semisupervised AG

- Use labeled data to extract counts of different rules and subtrees
- Labels must be compatible with the grammar
- Full bracketing is not required

Example Input:

(Morph s i n g) (Morph i n g)

Question 3

Regarding the model for morphological segmentation using Adaptor Grammars: Is it possible to use a small labeled set for both selecting a morphological template as in AG Select and for gathering counts from labeled segmentations as in semi-supervised AG?



M1 M2 M1 M21 M22 M11 M12 M2 M11 M12 M21 M22

- salt_iness \rightarrow salt_i_ness \rightarrow \rightarrow sal_t_iness \rightarrow

 - sal_t_i_ness



M1 M2 M1 M21 M22 M11 M12 M2 M11 M12 M21 M22 → salt_iness
→ salt_i_ness
→ sal_t_iness
→ sal_t_i_ness



M1 M2 M1 M21 M22 M11 M12 M2 M11 M12 M21 M22 → salt_iness
→ salt_i_ness
→ sal_t_iness
→ sal_t_i_ness
→ sal_t_i_ness



M1 M2 M1 M21 M22 **M11 M12 M2** M11 M12 M21 M22 → salt_iness
→ salt_i_ness
→ sal_t_iness
→ sal_t_i_ness



M1 M2 M1 M21 M22 M11 M12 M2 M11 M12 M21 M22 → salt_iness
→ salt_i_ness
→ sal_t_iness
→ sal_t_i_ness
→ sal_t_i_ness



M1 M2 M1 M21 M22 M11 M12 M2 M11 M12 M21 M22 → salt_iness
→ salt_i_ness
→ sal_t_iness
→ sal_t_i_ness
→ sal_t_i_ness

Feature-based similarity function



Distance-dependent Chinese restaurant process



 $P(\text{stepped} \rightarrow \text{played}) \propto e^{w^T f(\text{stepped, played})}$ $P(\text{table} \rightarrow \text{table}) \propto \alpha \qquad \qquad \text{pedantic}$

