Minimally-Supervised Morphological Segmentation Using Adaptor Grammars

Kairit Sirts and Sharon Goldwater

TALLINNA TEHNIKAÜLIKOOL / TALLINN UNIVERSITY OF TECHNOLOGY
INSTITUTE OF CYBERNETICS
THE UNIVERSITY of EDINBURGH informatics
ilcc Institute for Language, Cognition and Computation

Key Results

Two successful methods for using Adaptor Grammars for semisupervised learning:

- Semisupervised Adaptor Grammars
- AG Select

Showing how Adaptor Grammars can be used for inductive learning:

- Scales the AG for large data sets.
- On average no loss in learning accuracy.

Adaptor Grammars (Johnson et al., 2007)

Framework for specifying nonparametric Bayesian models to learn latent tree structures from a corpus of strings.

Two components of Adaptor Grammars:

- Base distribution, which is a PCFG.
- Pitman-Yor Process adaptor that "adapts" the probabilities assigned to individual subtrees under the PCFC model.

Example of a morphology learning grammars:

Word → Morph$^+$
Morph → SubMorph$^+$
SubMorph → Char$^+$

Underlined non-terminals are adapted
$^+$ abbreviates the recursive rules

Semisupervised AG

- Use labeled data to extract counts of different rules and subtrees.
- Labeled data must provide a consistent bracketing (no overlapping spans).
- Labels of the spans must be compatible with the grammar.
- Full bracketing of the labeled data is not required – the spans not specified will be induced by the AG.

Example Input (SubMorph structures will be induced)

(Morph s t a n d a r d) (Morph i z e) (Morph s)

AG Select

- Automatically identifies the best grammar for each language.
- Uses *metagrammar* to identify all possible morpheme borders in the words.
- Then generates all possible *morphological templates* for that metagrammar
- Finally uses a small amount of labeled data to select the most plausible template for each language.

Metagrammar

Metagrammar is a very general binary-branching grammar, so that:

- non-terminals near root should have high precision;
- non-terminals near leaves should have high recall.

Example grammar of depth 2:

Word → M1
Word → M1 M2
M1 → M11
M1 → M11 M12
M2 → M21
M2 → M21 M22
M11 → Char$^+$
M12 → Char$^+$
M21 → Char$^+$
M22 → Char$^+$

We use a grammar of depth 4, allowing max 16 segments per word.

Templates

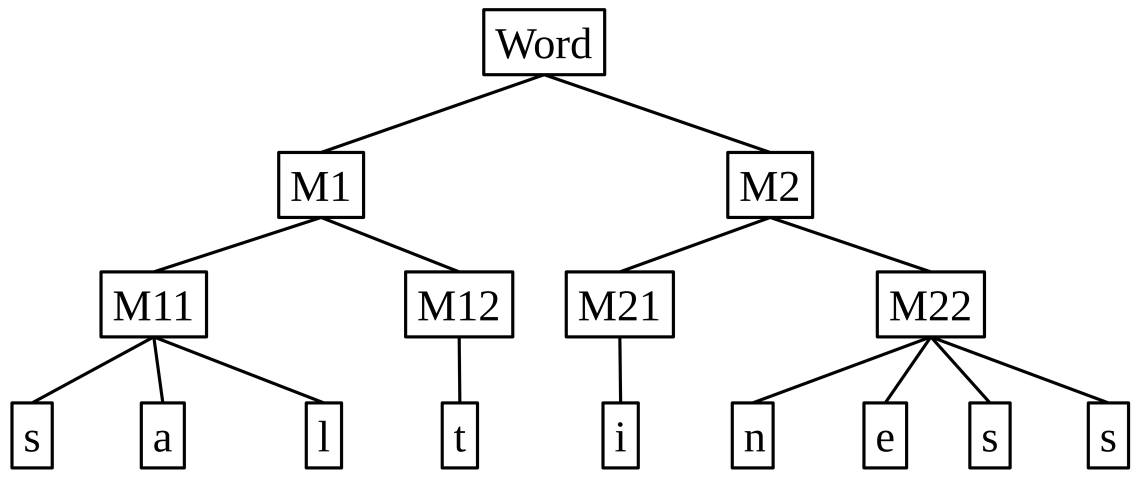
Morphological template:

-Used to select a segmentation of the word from the metagrammar parse tree.

- An ordered sequence of non-terminals whose yields span the full word.

Example:

- Parse tree with depth 2 metagrammar for the word **'saltiness'**



- All possible templates with corresponding segmentations:

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

Inductive Learning

1) Run the AG sampler on data set of feasible size.
2) Extract the posterior grammar (where rules include cached trees).
3) Use with standard parser to decode new data.

Results

Development set:

- Inductive learning performs as well as transductive learning.
- Semisupervised methods perform better than unsupervised baselines.

Model	Transductive Learning				Inductive Learning			
	Eng	**Est**	**Fin**	**Avg**	**Eng**	**Est**	**Fin**	**Avg**
AG unsup	66.2	66.9	60.5	64.5	66.1	67.5	61.6	65.1
Morfessor	69.5	55.7	65.0	63.4	68.9	51.1	63.5	64.5
AG ssv	70.0	67.5	**71.8**	69.8	**70.5**	67.2	**70.0**	**69.2**
AG Select	**71.9**	**68.5**	70.2	**70.2**	69.8	**68.8**	67.5	68.7

Segment border F-score

Test set:

- 600K – 2.9M words, so only trained inductively.
- Results are comparable to semisupervised Morfessor (Kohonen et al., 2010).

Model	F-score				EMMA			
	Eng	**Fin**	**Tur**	**Avg**	**Eng**	**Fin**	**Tur**	**Avg**
Morfessor	65.7	60.8	60.1	62.2	76.5	59.6	47.0	61.0
Morfessor ssv	**77.8**	**71.7**	**68.9**	**72.8**	80.6	62.1	**49.9**	64.2
AG ssv	70.3	64.9	58.2	64.5	75.9	61.3	46.1	61.1
AG Select	74.4	70.0	61.4	68.6	**81.3**	**64.0**	47.5	**64.3**

Summary

- Best average results on development set.
- Competitive results on test set.
- Learning latent submorph structures improves the overall results.

European Union European Social Fund
Investing in your future
DoRa