Unsupervised morphological segmentation and clustering with document boundaries

Taesun Moon Katrin Erk and Jason Baldridge

Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA

Empirical Methods in Natural Language Processing 2009

Introduction

Morphology acquisition

Morphology acquisition involves one or more of ...

- Segmentation of a word into constituent morphemes
  - inflectional: *morphemes* = *morpheme* + *s*
  - derivational: *segmentation* = *segment* + *ation*
  - indiscriminate: *morphemes* = *morph* + *eme* + *s*
- Clustering of words which are morphological variants
  cluster, clusters, clustered, clustering
- Generation of unobserved, inflected/derived word forms
  morpheme → *morphemes*

Introduction

Goals

Aid language documentation

- Documentation of endangered languages before they disappear
- Analysis of language data: typically by human annotators
- Aim: aid analysis using unsupervised machine learning
- Morphological preprocessing important part of producing Interlinearized Glossed Text

Use on data from endangered languages

- Allow use out of the box
- Minimize number of parameters
- Work with small amounts of data

Introduction

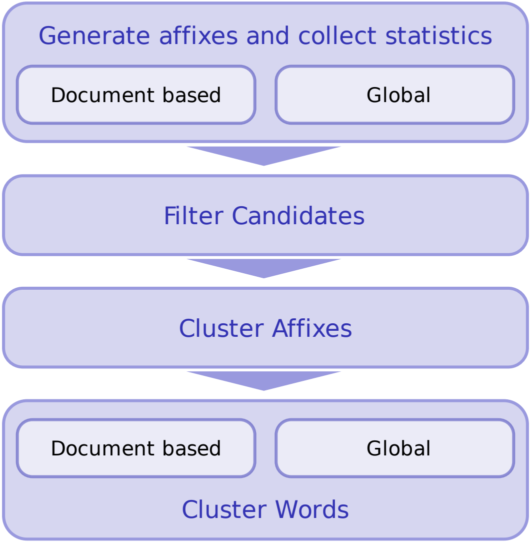
Core ideas

The core ideas of the model are . . .

- filter affixes by significant co-occurrence
- use document boundaries to eliminate noise

Model

Overview



Model

Stage I. Candidate Generation

- Build a trie from the lexicon of a document/all documents
- Split word into stem and affixes if paths after a branch are shorter than the path from the root to the branch
- Collect counts and pairwise counts for affixes

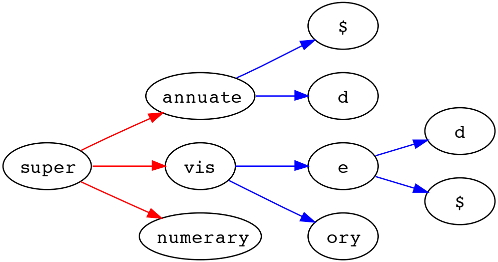


Figure: → neutral edges, → edges to affixes

Affixes (counts)

\$ (2), *s* (1), *d* (2), *ed* (1), *ory* (1)

Pairs (counts)

\$/d (2), *ory/e* (1), *ory/ed* (1), *e/ed* (1)

Model

Stage II. Candidate Filtering

Filtering rule

Only retain affix pairs which are significantly correlated under χ^2 test.

Sample counts: Doc

	ed	∼ed
ing	10273	21853
∼ing	27120	4119332

Table: χ^2=352678

	le	∼le
s	122	132945
∼s	936	4044575

Table: χ^2=239.132

Sample Counts: Global

	ed	∼ed
ing	2651	1310
∼ing	1490	150848

Table: χ^2=65101.6

	le	∼le
s	20	12073
∼s	198	144008

Table: χ^2=0.631($p = 0.427$)

Model

Stage III & IV

Stage III. Affix clustering

- Bottom up, minimum distance clustering
- Cluster membership is not exclusive and thus clusters are *not disjoint*

Stage IV. Word clustering

Cluster words iff

- the words occurred in the same document / global lexicon
- they have a shared path longer than some length in a trie defined for the document / global lexicon
- the affixes for these words belong to a cluster induced in stage iii.

Data

Training data

- two languages: English and Uspanteko
- for English, two data sets from NYTimes
  - one large (9M tokens), one small (187K tokens)
  - to simulate effect of small data sizes
- Uspanteko: Mayan language of K'ichee' branch with approx. 1320 speakers
- for Uspanteko, an even smaller data set (50K words)

English gold data

evaluate on the *inflectional* morphology portion of CELEX.

Uspanteko gold data

- use gold data from documentation project
- manually evaluate subsample of output

Evaluation

Metric

Basic counts

- Calculate numbers for correct ($\mathcal{C}$), inserted ($\mathcal{I}$) and deleted ($\mathcal{D}$) words.
- Take into account overlapping clusters
- Modification of Schone & Jurafsky (2001)

Scoring formula

Calculate precision (P), recall (R) and f-score (F):

$$
\begin{aligned}
P &= \mathcal{C}/(\mathcal{C}+\mathcal{I}) \\
R &= \mathcal{C}/(\mathcal{C}+\mathcal{D}) \\
F &= (2PR)/(P+R)
\end{aligned}
$$

Evaluation

Results: English

	MINI-NYT			NYT		
	P	R	F	P	R	F
LINGUISTICA	64.30	**93.34**	76.15	47.50	**88.33**	61.77
MORFESSOR	45.2	87.8	59.7	63.6	69.2	66.3
Cand-D +*Clust-G*	69.41	91.42	78.91	46.00	79.81	58.36
Cand-D +*Clust-D*	83.47	80.36	81.89	59.02	74.50	65.86
Cand-G +*Clust-G*	73.44	88.72	80.36	61.81	82.98	70.85
Cand-G +*Clust-D*	**88.34**	77.95	**82.82**	**77.71**	70.24	**73.79**

Table: Benchmarks performed with LINGUISTICA (Goldsmith, 2001) and MORFESSOR (Creutz and Lagus, 2007). (*Cand* = candidate generation; *Clust* = clustering; *D* = document-wise; *G* = global)

Evaluation

Results: Uspanteko (machine evaluation)

	P	R	F
Cand-G + *Clust-D*	**95.42**	47.89	63.78
Cand-G + *Clust-G*	92.03	50.01	**64.80**
LINGUISTICA	81.14	47.60	60.00
LINGUISTICA	84.15	52.00	64.28
MORFESSOR	28.12	**62.28**	38.75

Table: *Cand* = candidate generation; *Clust* = clustering; *D* = document-wise; *G* = global

Evaluation

Results: Uspanteko (expert evaluation)

	Acc.	FAcc.	Avg. Sz.
Cand-G + *Clust-G*	98.5	79.0	2.94
LINGUISTICA	96.0	85.0	2.64
MORFESSOR	85.3	55.0	4.8

Table: Human expert evaluated accuracy (Acc.), full cluster accuracy (FAcc.) and average cluster size in words (Avg. Sz.). Conducted on 100 non-singleton cluster subsamples. Full cluster accuracy is the number of clusters with no errors divided by subsample size (100)

Discussion I

Interaction of affix criterion and tries

- Global candidate generation more effective in filtering out spurious forms
- only long words generate candidates in global candidate generation
- chance of morphologically unrelated but orthographically similar short words coöccurring in same document increases with data size
- morphologically unrelated but orthographically similar words do generate candidates in global candidate generation but counts are suppressed

Discussion II

Summary

- Document clustering is effective in filtering out spurious members
- Document candidate generation enhances recall for small data sets.
- Model outperforms LINGUISTICA and MORFESSOR in terms of *f*-score and precision in all experiments.
- Model is simple, intuitive and flexible

Discussion III

Future work

- Approach not suited for languages with more complex morphology, e.g. agglutinative languages
- Performance deteriorates as size of data increases
  - perhaps phenomenon restricted to languages with relatively impoverished morphological inventory
  - similar results observed for English with LINGUISTICA here and MORFESSOR in Creutz and Lagus (2005).
  - approach seems feasible even with limited data for such languages