UNSUPERVISED MORPHOLOGICAL SEGMENTATION & CLUSTERING
OVERVIEW

- Introduction
- **Morphological Segmentation** (Creutz & Lagus 2005)
  - Aims
  - Models
  - Evaluation
  - Results
- **Affix Clustering** (Moon et al 2009)
  - Idea
  - Model
  - Results
- Conclusion
WHAT ARE WE DOING?

- Morpheme Segmentation
  - Morphemes = smallest meaning-bearing units
  - = smallest elements of syntax
- Meaning vs. Form
- Composition vs. Perturbation

reads = read + s
machines = machine + s
translation = translate + ion
goalkeeper = goal + keeper
joystick = joy + stick
WHAT ARE WE DOING?

- **Stem** vs. **Affixes** (Prefixes + Suffixes)
- **Inflectional** vs. **Derivational**
- **Affix Clustering**

Talk
Talks = talk + s
Talked = talk + ed
Talking = talk + ing

Teach
Teachers = teach + er + s
Teaches = teach + es
Teaching = teach + ing
WHY ARE WE DOING IT?

- important information

- especially for highly inflected languages
  (like Turkish, Finnish, Nahuatl, Japanese → agglutinative languages)

- used in other CL applications
  (language production, speech recognition, machine translation etc.)
„algorithm for the unsupervised learning […] of a simple morphology of a natural language“

Unsupervised morpheme segmentation with hierarchical representation

English and Finnish
AIMS

- Most accurate segmentation possible
- Learn representation of the language in the data + store it in a lexicon
- Based several models: Linguistica, Morfessor Baseline, Morfessor ML, Morfessor MAP
**BASELINE**

- **Morfessor Baseline** Algorithm (Creutz&Langus)

- Similar to some unsupervised word segmentation algorithms
- Construct lexicon of morphs
- Each word can be constructed out of those morphs
- **AIM:** find optimal + concise segmentation and lexicon
- **PROBLEM:** frequent words stored as a whole - rare words excessively split + stored in part
  
  no representation of a morph's inner structure
- *Linguistica* (Goldsmith 2001)

- Splits word into stem + one (empty) prefix / affix

- **ADVANTAGE:** Modeling of simple word-internal syntax (morphotactics – rules on ordering of morphemes) – grouping sets of stems & suffixes into inflectional paradigms

- **DISADVANTAGE:** handles highly inflecting + compounding languages poorly (alternating stems + affixes)
**IMPROVED MODEL**

- **Morfessor Categories-ML** (Creutz&Lagus)
- Reanalyzes segmentation of *Morphessor Baseline*
- Maximum Likelihood Model
- Words represented as HMMs
- Stems, prefixes + suffixes can alternate (with some restrictions)
- „noise“ category
- → split words whose morphs are present in the lexicon
- → join „noise“ morphs with their neighbours to form proper morphs
- **CRITICISM:** too ad hoc + information on word frequency is lost

hidden states: categories (SUFF, PRE, …)

observable states: morphs
NEW MODEL

- **Morfessor Categories-MAP** (Creutz & Lagus)
- Induces binary hierarchical lexicon
- Retains inner structure of words → morphs represented as concatenation of (sub)morphs of the lexicon
- Word frequency (own entry vs. Split into morphs)
- Prefix – Stem – Suffix – Non-morpheme

![Diagram showing morphological structure of words with prefixes, stems, and suffixes.](image-url)
- Maximum a posteriori framework
- Words represented as HMMs
- Desired level of segmentation: „finest resolution that does not contain non-morphemes“
SEARCH ALGORITHM (GREEDY SEARCH)

Initialisation of segmentation

Splitting of morphs

Joining of morphs

Splitting of morphs

Resegmentation of corpus + re-estimation of probabilities

Expansion to finest resolution

Representativeness + ness
stem + SUFF

[Re+[present+ativ]]+[n+ess]
PRE+stem+SUFF+non+SUFF

[Re+[present+ativ]]+ness
PRE+stem+SUFF+SUFF

[Re+[[pre+sent]+ativ]]+ness
PRE+non+stem+SUFF+SUFF

[Re+[[pre+sent]+ativ]]+ness
PRE+non+stem+SUFF+SUFF

[Re+[present+ativ]]+ness
PRE+stem+SUFF+SUFF
AIM: Finding optimal lexicon + segmentation

Maximum a posteriori estimate to be maximized:

\[
\text{arg max}_{\text{lexicon}} P(\text{lexicon} | \text{corpus}) = \text{arg max}_{\text{lexicon}} P(\text{corpus} | \text{lexicon}) \cdot P(\text{lexicon})
\]

\[
P(\text{corpus} | \text{lexicon}) = \prod_{j=1}^{W} [P(C_{j1} | C_{j0}) \prod_{k=1}^{n_j} P(\mu_{jk} | C_{jk}) \cdot P(C_{j(k+1)} | C_{jk})]
\]

Morph emission probability

\[
P(\text{lexicon}) = M! \cdot \prod_{i=1}^{M} [P(\text{meaning}(\mu_i)) \cdot P(\text{form}(\mu_i))]
\]
- **Morph Emission Probabilities**
- $\rightarrow$ probability that morph is emitted by the category
- Depend on frequency of morph in training data
- *Prefix-/Suffix-Likeness* (right+left perplexity)
- *Stem-Likeness* (length)
- Non-morpheme probability

\[
P(\mu_{jk} \mid C_{jk}) = \frac{P(C_{jk} \mid \mu_{jk}) \cdot P(\mu_{jk})}{P(C_{jk})} \quad (8)
\]

\[
= \frac{P(C_{jk} \mid \mu_{jk}) \cdot P(\mu_{jk})}{\sum_{\forall \mu_{j'k'}} P(C_{jk} \mid \mu_{j'k'}) \cdot P(\mu_{j'k'})}.
\]
EVALUATION

Finnish Data
- Prose + news text
- Finnish IT Centre of Science
- Finnish National News Agency

English Data
- Prose + news + scientific text
- Gutenberg Project
- Gigaword Corpus
- Brown Corpus

Goldstandard
- Hutmegas
  - Linguistic morpheme segmentations
  - 1.4 million Finnish
  - 120,000 English word forms

Evaluation on
10,000, 50,000, 250,000, 12/16 million words
RESULTS
Simple model without heuristics /thresholds /trained parameters

- **Word segmentation** - constrain candidate stems + affixes by document boundaries

- **Cluster affixes** of certain stems → morphologically related words

- **USE**: interlinearised glossed texts for LRL

- English + Uspanteko
IDEA

- two words in the same document are very similar in orthography → likely to be related morphologically
- use document boundaries to filter out noise
- constrain potential membership of word clusters

Our intuition comes from an observation by Yarowsky (1995) regarding multiple tokens of words in documents. He tabulates the applicability of using document boundaries to disambiguate word senses, which measures how often a given word occurs more than twice in the same document. For ten potentially ambiguous words, he counts how often they occur more than once in some document and finds that if the words do occur, they do so multiple times in 50.1% of these documents, on average. His counts ignored morphological variation, and it is likely the applicability measure would have increased considerably: if a content word is used more than once in some text, it is likely to be repeated in different syntactic contexts, requiring the word to be inflected or to be derived for a different part-of-speech category.³

He suddenly drew a sharp sword... The documentation of...
Conflation set:

„Set of word types that are related through either inflectional or derivational morphology“
CANDIDATE TRIE

Stems ← trunks

branches → affixes
Conflation set:

“Set of word types that are related through either inflectional or derivational morphology”

X2 testing: Correlation betw. Affixes

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$\sim b_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_2$</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
</tr>
<tr>
<td>$\sim b_2$</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum_{x \in C} \frac{(O_{xy} - E_{xy})^2}{E_{xy}} \]

\[ \begin{array}{c|c|c|c|c} \text{ed} & \sim \text{ed} & \text{le} & \sim \text{le} \\ \hline \text{ing} & 10273 & 21853 & 122 & 132945 \\ \sim \text{ing} & 27120 & 4119332 & 936 & 4044575 \\ \end{array} \]
## RESULTS

<table>
<thead>
<tr>
<th></th>
<th>MINI-NYT</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td><strong>LINGUISTICA</strong></td>
<td>64.30</td>
<td>93.34</td>
</tr>
<tr>
<td><strong>MORFESSION</strong></td>
<td>45.2</td>
<td>87.8</td>
</tr>
<tr>
<td><strong>CandGen-D + Clust-G</strong></td>
<td>69.41</td>
<td>91.42</td>
</tr>
<tr>
<td><strong>CandGen-D + Clust-D</strong></td>
<td>83.47</td>
<td>80.36</td>
</tr>
<tr>
<td><strong>CandGen-G + Clust-G</strong></td>
<td>73.44</td>
<td>88.72</td>
</tr>
<tr>
<td><strong>CandGen-G + Clust-D</strong></td>
<td><strong>88.34</strong></td>
<td>77.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>FAcc.</th>
<th>Avg. Sz.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ca-G + Cl-G</strong></td>
<td>98.5</td>
<td>79.0</td>
<td>2.94</td>
</tr>
<tr>
<td><strong>LINGUISTICA</strong></td>
<td>96.0</td>
<td>85.0</td>
<td>2.64</td>
</tr>
<tr>
<td><strong>MORFESSION</strong></td>
<td>85.3</td>
<td>55.0</td>
<td>4.8</td>
</tr>
</tbody>
</table>
Thank you for your attention!