UNGRADE: UNsupervised GRAph DEcomposition

Bruno Golénia, Sebastian Spiegler, Peter Flach

University of Bristol, Department of Computer Science

September 30th, 2009
Outline

Introduction

UNGRADE algorithm
   Stem extraction
   Graph structure
   Scoring function for merging process
   Stopping criterion for merging process

Experiments

Conclusions and future work
Outline

**Introduction**

**UNGRADE algorithm**
- Stem extraction
- Graph structure
- Scoring function for merging process
- Stopping criterion for merging process

**Experiments**

**Conclusions and future work**
Introduction

Objective

- Unsupervised word decomposition

Assumptions

- Word = prefix sequence + stem + suffix sequence
- No restrictions on the number of prefixes and suffixes
- Each word has one stem
UNGRADE: Three steps algorithm

- **Stem extraction** using letter window
- **Graph structure** for finding prefixes and suffixes
- **Aggregation** of prefixes, stems and suffixes
Outline

Introduction

UNGRADE algorithm
- Stem extraction
- Graph structure
- Scoring function for merging process
- Stopping criterion for merging process

Experiments

Conclusions and future work
Minimum description length window score

- **Window** with a left boundary \( l_{\text{win}} \) and an right boundary \( r_{\text{win}} \)
  \[ \text{win} = (l_{\text{win}}, r_{\text{win}}) \]
- **Minimum description length window score** given word \( w \) and window \( \text{win} \)
  \[ \text{MDLWS}(\text{win}, w) = \log_2(r_{\text{win}} - l_{\text{win}} + 1) + \log_2(\text{npss}(w, l_{\text{win}}, r_{\text{win}})) \]
  \( \text{npss} \) denotes the n-gram probability of window \( \text{win} \) in word \( w \)
- **Window characteristics**:
  operators: shift, increase, decrease
  convergence: at optimum for **minimum description length window score**
- **Examples**: gearb||eitet → ge|arbeit|et, gela||ufen → ge|lauf|en
Graph structure $\rightarrow$ Morpheme graph

- node = letter $\xrightarrow{\text{merging process}}$ node = morpheme
- Use bottom-up approach to create morphemes
- Merge nodes using position-independent n-gram statistics
- Stop merging according to Bayesian Information Criterion and Jensen-Shannon divergence
Scoring function for merging node pairs

- Merging nodes requires function to score each pair of nodes in the graph
- Our merging function *Morph*Lift is based on lift of association rules [Brin et al. 97] and defined as

\[
Morph\_Lift(m_1, m_2) = \frac{f_{1,2}}{f_1 + f_2}
\]

for morpheme pair \((m_1, m_2)\) with \(f_i\) as frequency of morpheme \(m_i\)
- Pair of morphemes which maximises *Morph*Lift is used for merging
Jensen-Shannon divergence

Decrease in entropy between concatenated and individual morphemes for two morphemes $m_1$ and $m_2$ [Li 01]:

$$D_{JS}(m_1, m_2) = H(m_1 \cdot m_2) - \frac{L_{m_1}H(m_1) + L_{m_2}H(m_2)}{N}$$

where $H(m) = -p(m) \log_2 p(m)$, and $N = \sum_m \text{Freq}(m)$.

Stopping criterion

Requires that $\Delta BIC < 0$ which translates to:

$$\max_{m_1, m_2} D_{JS}(m_1, m_2) \leq 2 \log_2 N$$
Outline

Introduction

UNGRADE algorithm
- Stem extraction
- Graph structure
- Scoring function for merging process
- Stopping criterion for merging process

Experiments

Conclusions and future work
### Experimental Results

#### Morpho Challenge 2009

<table>
<thead>
<tr>
<th>Language</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (non-vowelized)</td>
<td>.8348</td>
<td>.1595</td>
<td>.2678</td>
</tr>
<tr>
<td>Arabic (vowelized)</td>
<td>.7215</td>
<td>.4361</td>
<td>.5436</td>
</tr>
<tr>
<td>English</td>
<td>.2829</td>
<td>.5174</td>
<td>.3658</td>
</tr>
<tr>
<td>Finnish</td>
<td>.4078</td>
<td>.3302</td>
<td>.3649</td>
</tr>
<tr>
<td>German</td>
<td>.3902</td>
<td>.2925</td>
<td>.3344</td>
</tr>
<tr>
<td>Turkish</td>
<td>.4667</td>
<td>.3016</td>
<td>.3664</td>
</tr>
</tbody>
</table>

---

**Results Morpho Challenge 2009**

- **Precision**
- **Recall**
- **F-Measure**
Outline

Introduction

UNGRADE algorithm
  Stem extraction
  Graph structure
  Scoring function for merging process
  Stopping criterion for merging process

Experiments

Conclusions and future work
Conclusions and future work

- Good results for a simple approach
- Similar F-measure for English, German, Turkish and Finnish
- Best results for vowelized Arabic
- High performance for languages with long words and high number of morphemes
- Use a Committee approach with selection of segmentation through description length
Thank you for your attention!