UNGRADE:
UNsupervised GRAph DEcomposition

Bruno Golénia, Sebastian Spiegler, Peter Flach

University of Bristol, Department of Computer Science

September 30th, 2009

Outline

Outline

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

Introduction

UNGRADE: Three steps algorithm

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

Outline

UNGRADE: Stem extraction

Minimum description length window score

- **Window** with a left boundary l_{win} and an right boundary r_{win}

  $win = (l_{win}, r_{win})$
- **Minimum description length window score** given word w and window win

  $MDLWS(win, w) = log_2(r_{win} - l_{win} + 1) + log_2(npss(w, l_{win}, r_{win}))$

  $npss$ denotes the n-gram probability of window win in word w
- **Window characteristics**:

  operators: shift, increase, decrease

  convergence: at optimum for **minimum description length window score**
- **Examples**: gearb||eitet $\longrightarrow$ ge|arbeit|et, gela||ufen $\longrightarrow$ ge|lauf|en

UNGRADE: Graph structure

Graph structure $\longrightarrow$ 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

UNGRADE: Scoring function for merging process

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

UNGRADE: Stopping criterion for merging process

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 $N = \sum_m Freq(m)$.

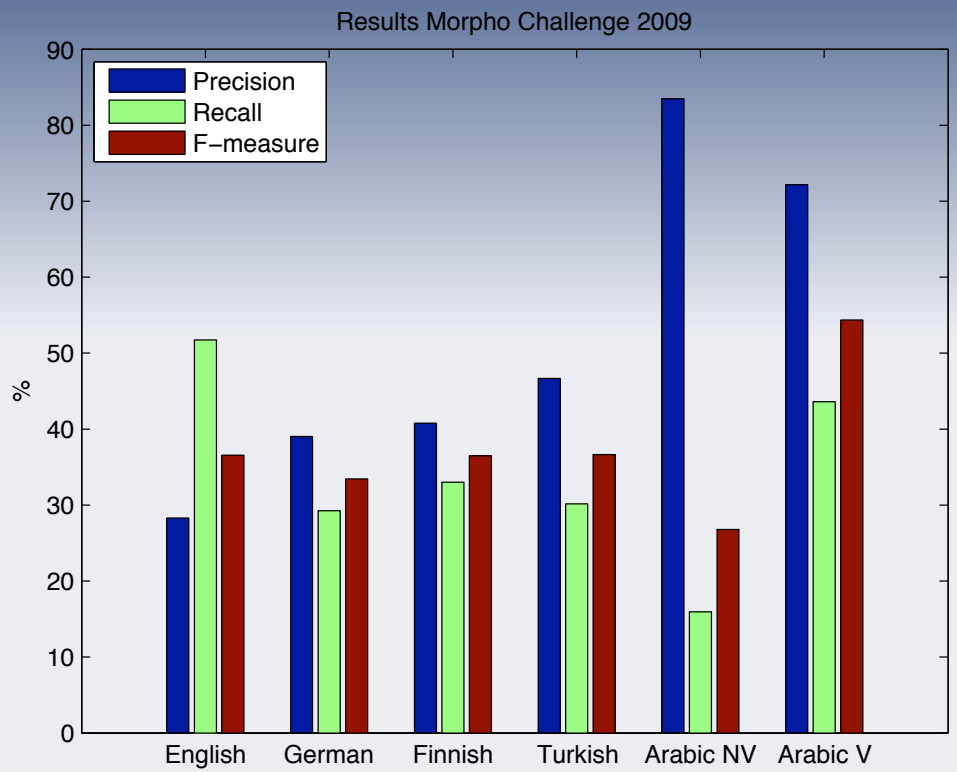
Stopping criterion

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

Experiments



Language	Precision	Recall	F-Measure
Arabic (non-vowelized)	.8348	.1595	.2678
Arabic (vowelized)	.7215	.4361	**.5436**
English	.2829	.5174	.3658
Finnish	.4078	.3302	.3649
German	.3902	.2925	.3344
Turkish	.4667	.3016	.3664

Outline

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!