Putting Suffix-Tree-Stemming to Work

Benno Stein
Bauhaus University Weimar

Martin Potthast
Paderborn University
Index terms

Text with markups  [Reuters]:

<TEXT>  <TITLE>CHRYSLER> DEAL LEAVES UNCERTAINTY FOR AMC WORKERS</TITLE>  <AUTHOR> By Richard Walker, Reuters</AUTHOR>  <DATELINE> DETROIT, March 11 - </DATELINE><BODY>Chrysler Corp’s 1.5 billion dlr bid to takeover American Motors Corp; AMO> should help bolster the small automaker’s sales, but it leaves the future of its 19,000 employees in doubt, industry analysts say. It was "business as usual" yesterday at the American ...
chrysler deal leaves uncertainty for amc workers
by richard walker reuters detroit march 11
chrysler corp s 1.5 billion dlr bid to takeover
american motors corp should help bolster the
small automaker s sales but it leaves the future
of its 19,000 employees in doubt industry
analysts say it was business as usual yesterday
at the american
Index terms

Stop words emphasized:

chrysler deal leaves uncertainty for amc workers by richard walker reuters detroit march 11
chrysler corp s 1 5 billion dlr bid to takeover american motors corp should help bolster the
small automaker s sales but it leaves the future of its 19 000 employees in doubt industry
analysts say it was business as usual yesterday at the american
Index terms

After stemming:

chrysler deal leav uncertain amc work richard walk reut detroit takeover american motor help bols automak sal leav futur employ doubt industr analy business usual usual yesterday
Index terms

After stemming:

chrysler deal leav uncertain amc work richard walk reut detroit takeover american motor help bols automak sal leav futur employ doubt industr analy business usual yesterday

Stemming algorithms remove inflectional and morphological affixes.

connect  
connects  
connected  
connecting  
connection
Index terms

After stemming:

chrysler deal leav uncertain amc work richard 
walk reut detroit takeover american motor help 
bols automak sal leav futur employ doubt industr 
analy business usual yesterday

Stemming algorithms remove inflectional and morphological affixes.

connect connects
connected
connecting
collection

+ make text operations less dependent on special word forms
+ reduce the dictionary size

– may merge words that have very different meanings
– discard possibly useful information about language use
Index terms

- Boolean model
- Fuzzy set model
- Vector space model
- Probabilistic model (BIR, NBIR, Poisson, etc.)
- Algebraic model
- Inference network model
- Generative language model (statistical language model)
- Suffix model
- Text structure model
- Word class model
- Direct usage of document terms
- Hidden variables and concepts
- Information on structure
- Special linguistic features
- Linguistic theory

Retrieval model ~ document model
Stemming Approaches

1. Table lookup.
   To each stem all flections are stored in a hash table.
   Problem: memory size (consider client-side applications)

2. Successor variety analysis.
   Morpheme boundaries are found by statistical analyses.
   Problem: parameter settings, runtime

3. Affix elimination.
   Rule-based replacement of prefixes and suffixes;
   the most commonly used approach.

Principle: *iterative longest match stemming*

(a) Removal of the match resulting from the longest precondition.
(b) Exhaustive application of the first step.
(c) Repair of irregularities.
## Stemming Approaches

### Affix Elimination under Porter

<table>
<thead>
<tr>
<th>Rule type</th>
<th>Condition</th>
<th>Suffix</th>
<th>Replacement</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Null</td>
<td>sses</td>
<td>ss</td>
<td>caresses → caress</td>
</tr>
<tr>
<td>1a</td>
<td>Null</td>
<td>ies</td>
<td>i</td>
<td>ponies → poni</td>
</tr>
<tr>
<td>1b</td>
<td>(m&gt;0)</td>
<td>eed</td>
<td>ee</td>
<td>feed → feed</td>
</tr>
<tr>
<td>1b</td>
<td>(<em>v</em>)</td>
<td>ed</td>
<td>ε</td>
<td>plastered → plaster</td>
</tr>
<tr>
<td>1b</td>
<td>(<em>v</em>)</td>
<td>ing</td>
<td>ε</td>
<td>motoring → motor</td>
</tr>
<tr>
<td>1c</td>
<td>(<em>v</em>)</td>
<td>y</td>
<td>i</td>
<td>happy → happi</td>
</tr>
<tr>
<td>2</td>
<td>(m&gt;0)</td>
<td>biliti</td>
<td>ble</td>
<td>sensibiliti → sensible</td>
</tr>
</tbody>
</table>

(m>x) number of vocal-consonant-sequences exceeds x
(∗S) stem ends with letter S
(∗v∗) stem contains vocal
(∗o) stem ends with cvc where second consonant c ∉ {W, X, Y}
(∗d) stem ends with two identical consonants
Stemming Approaches

Affix Elimination under Porter: Weaknesses

- difficult to modify:
  effects of new rules are barely to anticipate

- subject to over-generalization:
  policy/police university/universe
  organization/organ

- several definite generalizations are not covered:
  European/Europe matrices/matrix
  machine/machinery

- generates stem that are hard to be interpreted:
  iteration/iter general/gener
Stemming Approaches

Successor Variety Analysis: Interesting Aspects

- The idea of *corpus-specific stemming*. Corpus dependency is an advantage, if the corpus has a strong topic or application bias.

- The idea of *language independence*. Language independence is essential for multilingual documents or if the language cannot be determined.

<table>
<thead>
<tr>
<th>Stemming approach</th>
<th>Corpus dependency</th>
<th>Language independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affix elimination</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Variety analysis</td>
<td>yes</td>
<td>little</td>
</tr>
</tbody>
</table>
Stemming Approaches

Successor Variety Analysis: Realization

Suffix tree at letter level:
Stemming Approaches

Successor Variety Analysis: Realization

Suffix tree at letter level:

Suffix tree at word level:
Stemming Approaches

Successor Variety Analysis: Realization

Suffix tree at letter level:

- `nect` 1
- `con` 2
- `ing` 3
- `s` 1
- `ed` 1

Suffix tree at word level:

- `father plays chess` 0
- `boy plays chess too` 1
- `chess` 2
- `plays chess too` 2
- `too` 1

How to find good candidates for a stem?

- analysis of degree differences (depending on tree depth)
- cut-off method, complete word method, entropy method
Evaluation

Caution is advised ; )

- existing reports on the impact of stemming are contradictory
- employed analysis tool (among others): clustering

But what can be found?

1. improved document model
2. peculiarity of a clustering algorithm
3. . . .
Evaluation

Caution is advised ; )

- existing reports on the impact of stemming are contradictory
- employed analysis tool (among others): clustering

But what can be found?

1. improved document model
2. peculiarity of a clustering algorithm
3. . . .

A cluster algorithm’s performance depends on various parameters. Different cluster algorithms behave differently sensitive to document model “improvements”.

Baseline? Interpretation? Objectivity? Generalizability?
Evaluation

Caution is advised ; )

An objective way to rank document models is to compare their ability to capture the intrinsic similarity relations of a collection $D$.

Basic idea:

1. construct a similarity graph, $G = \langle V, E, w \rangle$
2. measure its conformance to a reference classification
3. analyze improvement/decline under new document model
Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- $G$ is called sparse [dense] if $|E| = O(|V|) \ [O(|V|^2)]$
- the density $\theta$ computes from the equation $|E| = |V|^\theta$
Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- $G$ is called sparse [dense] if $|E| = O(|V|)$ [$O(|V|^2)$]
- the density $\theta$ computes from the equation $|E| = |V|^\theta$

- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G') = |V|^\theta \iff \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

Using $\theta$ we assess the density of an induced subgraph $G_i$ of $G$. 
Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- $G$ is called sparse [dense] if $|E| = O(|V|)$ [$O(|V|^2)$]
- the density $\theta$ computes from the equation $|E| = |V|^\theta$

- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G') = |V|^\theta \iff \theta = \frac{\ln (w(G))}{\ln (|V|)}$$

Using $\theta$ we assess the density of an induced subgraph $G_i$ of $G$.

- a categorization $C = \{C_1, \ldots, C_k\}$ induces $k$ subgraphs $G_i$

$\Rightarrow$ expected density $\bar{\rho}(C) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^\theta}$
Expected Density $\bar{\rho}$

Understanding Expected Density

Embedding of a collection under a particular document model.
Expected Density $\bar{\rho}$

Understanding Expected Density

Embedding of a collection under a particular document model.

$\bar{\rho} > 1$  [$\bar{\rho} < 1$] if the cluster density is larger [smaller] than average.
Expected Density $\bar{\rho}$

Understanding Expected Density

Consider inter-cluster and intra-cluster similarities.
Expected Density $\bar{\rho}$

Understanding Expected Density

Consider inter-cluster and intra-cluster similarities.

Effect of a document model that *reinforces the structural characteristic* within a document collection.
The expected density $\bar{\rho}$ is a monotonically increasing function of the sample size.
The expected density $\bar{\rho}$ is a monotonically increasing function of the sample size.
The expected density \( \bar{\rho} \) is a monotonically increasing function of the sample size.
The expected density $\bar{\rho}$ is a monotonically increasing function of the sample size.
Collection: RCV1. Two documents $d_1, d_2$ are assigned to the same category if they share the top level category and the most specific category.
**Expected Density** $\bar{\rho}$

**Experiments: English Collection**

A note on reproducibility: meta information files that describe the compiled test collections are made available upon request.
Expected Density $\bar{\rho}$

Experiments: German Collection

Collection: Compilation of 26,000 documents from 20 German news groups.
Expected Density $\bar{\rho}$

Experiments: German Collection

![Graph showing expected density $\bar{\rho}$ for different stemming approaches. The graph includes lines for Stemming: without, Snowball, and Suffix tree, with increasing density as the sample size increases. There are 5 categories displayed on the graph.](image-url)
Expected Density $\bar{\rho}$

Experiments: German Collection

Stemming can reduce noise.
### Expected Density $\bar{\rho}$

#### Experiments: German Collection

Where successor variety works:

<table>
<thead>
<tr>
<th>Term</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>mechanism</td>
<td>mus, tisch, che, ch, tischen, men, tisches, ierung, chen</td>
</tr>
<tr>
<td>zusammen</td>
<td>leben, gang, h</td>
</tr>
<tr>
<td>zusammenbr</td>
<td>icht, uch, aut, echen</td>
</tr>
<tr>
<td>zusammenfass</td>
<td>en, ung, t, end</td>
</tr>
<tr>
<td>zusammengeme</td>
<td>faßt, baut, zählt, fasst</td>
</tr>
<tr>
<td>zusammengesetzt</td>
<td>en, $</td>
</tr>
<tr>
<td>zusammenh</td>
<td>ängen, ängt, änge</td>
</tr>
<tr>
<td>zusammenha</td>
<td>1ten, lt</td>
</tr>
<tr>
<td>zusammenhang</td>
<td>los, es, s, $</td>
</tr>
</tbody>
</table>
**Expected Density** $\bar{\rho}$

**Experiments: German Collection**

Where successor variety works:

<table>
<thead>
<tr>
<th>Word</th>
<th>Stemmed Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>mechanis</td>
<td>mus, tisch, che, ch, tischen, men,</td>
</tr>
<tr>
<td></td>
<td>tisches, ierung, chen</td>
</tr>
<tr>
<td>zusammen</td>
<td>leben, gang, h</td>
</tr>
<tr>
<td>zusammenbr</td>
<td>icht, uch, aut, echen</td>
</tr>
<tr>
<td>zusammenfass</td>
<td>en, ung, t, end</td>
</tr>
<tr>
<td>zusammenenge</td>
<td>faßt, baut, zählt, fasst</td>
</tr>
<tr>
<td>zusammengesetzt</td>
<td>en, $</td>
</tr>
<tr>
<td>zusammenh</td>
<td>ängen, ängt, änge</td>
</tr>
<tr>
<td>zusammenha</td>
<td>lten, lt</td>
</tr>
<tr>
<td>zusammenhang</td>
<td>los, es, s, $</td>
</tr>
</tbody>
</table>

and where it fails:

<table>
<thead>
<tr>
<th>Word</th>
<th>Stemmed Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>schwarz</td>
<td>arbeit, denker, schild, fahrer, en,</td>
</tr>
<tr>
<td></td>
<td>e, markt, maler, bader, hörer, radler, e, s</td>
</tr>
</tbody>
</table>
## Evaluation

### A Note on $F$-Measure Values

<table>
<thead>
<tr>
<th>Stemming approach</th>
<th>$F$-min (sample size 1000, 10 categories)</th>
<th>$F$-max</th>
<th>$F$-av.</th>
</tr>
</thead>
<tbody>
<tr>
<td>without—baseline—</td>
<td>—— baseline ——</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>Porter</td>
<td>-12%</td>
<td>11%</td>
<td>2%</td>
</tr>
<tr>
<td>suffix tree</td>
<td>-10%</td>
<td>10%</td>
<td>2%</td>
</tr>
</tbody>
</table>
Evaluation

A Note on $F$-Measure Values

<table>
<thead>
<tr>
<th>Stemming approach</th>
<th>$F$-min (sample size 1000, 10 categories)</th>
<th>$F$-max</th>
<th>$F$-av.</th>
</tr>
</thead>
<tbody>
<tr>
<td>without —baseline—</td>
<td>——baseline—</td>
<td>——baseline—</td>
<td>——baseline—</td>
</tr>
<tr>
<td>Porter</td>
<td>-12%</td>
<td>11%</td>
<td>2%</td>
</tr>
<tr>
<td>suffix tree</td>
<td>-10%</td>
<td>10%</td>
<td>2%</td>
</tr>
</tbody>
</table>

A Note on Runtime

- successor variety analysis with suffix trees
  in $O(n)$ [Ukkonen 1995], and
  in $O(n^2)$ and $\Theta(n \log(n))$ respectively [Giegerich et. al.]

- successor variety analysis with Pat trees
  in $O(n^2)$; $\Theta(n \log(n))$ may be assumed for short affixes
Summary

- Basis: document models with “visible” index terms
- Issue: selection, modification, enrichment of index terms
- Question: stemming without semantic background

Contribution

- efficient implementation of variational stemming with Patricia
- parameter optimization \( \Rightarrow \) significantly better than [Frakes 1992]
- comparison to Porter stemmer and Snowball stemmer
- algorithm-neutral evaluation method based on \( \bar{\rho} \)

Message

- the impact of stemming may be over-estimated
- generally accepted analysis methods are required
Summary

Related Work

- A similar approach can be applied to index construction. *variational* n-grams: use words (not letters) as tokens
- Issue: *collection-specific* document model
- Motto: “co-occurrence analysis versus Wordnet”
Summary

Related Work

- A similar approach can be applied to index construction. *variational* n-grams: use words (not letters) as tokens
- Issue: *collection-specific* document model
- Motto: “co-occurrence analysis versus Wordnet”

![Graph showing expected density ρ vs sample size]

**Additional concepts:** without, Wordnet, *n*-gram
References


Introduction
Stemming Approaches
Evaluation

Σ

GFKL'06  Mar. 8th, 2006
Stein/Potthast