Folktales As Classifiable Texts

Learning to Extract Folktale Keywords

Dolf Trieschnigg, Dong Nguyen and Mariët Theune
Once upon a time…

• There was a research institute in Amsterdam that wanted to collect folktales…
• Not only to study Dutch folklore, but also to document part of the Dutch oral tradition...
• They asked people from all over the Netherlands to collect stories in their surroundings
• How did they do that in a time without tablets, smartphones and laptops?
They got on their bike and used pencil and paper. Later they even used tape recorders...
- They stored these stories in large archives to collect dust and to be used by researchers.
In 1994 they started inserting the archives in a so-called database: The Dutch Folktale Database was born...
They employed students to digitize these paper stories, add metadata, and store them in the database.
• In 2004 the database became available online!
• So the Meertens Institute lived happily ever after?
• No, because still too many stories await archiving; adding metadata takes too much time.
• So they decided to study automatic keyword extraction.
Overview

- About the collection: The Dutch Folktale Database
- **Characteristics** of keywords in the DFDB
  - Statistics
  - How do the keywords relate to the story text?
  - Do annotators agree?
- **Automatic extraction** of keywords
  - Setup, systems & results
  - Which features to use?
- Conclusion
The Dutch Folktale Database

- Maintained by the Meertens Institute since 1994
- > 40,000 Dutch folktales, collected since the 19th century
- Subgenres
  - Fairy tales, legends, urban legends
  - jokes, riddles, personal narratives
- Languages
  - Dutch, Frisian, Old Dutch, Middle Dutch and many Dutch dialects
- Other metadata
  - Summary, keywords, story type, motifs
  - proper names, storyteller, location etc.
- Online since 2004: www.verhalenbank.nl
Keywords in the DFDB (1/2)
Keywords in the DFDB (2/2)

- Keyword assignment
  - Manual uncontrolled vocabulary indexing
  - Vaguely defined indexing task
  - Carried out by many different annotators
- Statistics (42k docs, 17k Dutch)
  - 15 assigned keywords on average, median 10
  - Mostly single words (90%)
  - 43k unique keywords
  - 65% of keywords appears literally in (Dutch) text
How do the keywords relate to the story text?

- Manual classification of 50 docs, 989 keywords

- Classes
  - Literal: 68%
  - Almost literal: 12%
  - Synonym: 5%
  - Hypernym: 2%
  - Typing error: <1%
  - Other (more abstract, etc.): 13%

- 80% can be (almost) literally linked to the text
Do annotators agree?

• Setup
  • 10 annotators (2 experienced), 5 stories each
  • Each story annotated by 2 annotators
  • Judge all story words:
    1) non-relevant; 2) relevant; 3) highly relevant
  • Determine inter-annotator agreement
• Results:
  • Substantial agreement on relevant keywords ($\kappa$: 0.62), only moderate agreement on highly relevant keywords ($\kappa$: 0.48)
  • Reasons for disagreement
    1) verbs and adjectives? 2) overlooked
    3) choice rather than both 4) lack of instructions
  • Experienced annotators indicate more relevant keyword and show higher average agreement
Automatic extraction

• Setup
  • Ranking task: rank most relevant words from text first
  • Evaluation: reproduce manual keyword list (IR metrics)
  • 17,000 documents, 10-fold cross-validation

• Systems
  • Baseline 1: TF-IDF (in training collection)
  • Baseline 2: TF-IDF-T (prefer seen keywords)
  • Learning to rank: linear ranking SVM
    • Features from word, document and collection context

• Results

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>P@5</th>
<th>P@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>0.260</td>
<td>0.394</td>
<td>0.317</td>
</tr>
<tr>
<td>TF-IDF-T</td>
<td>0.336</td>
<td>0.541</td>
<td>0.384</td>
</tr>
<tr>
<td>rank-SVM</td>
<td>0.399</td>
<td>0.631</td>
<td>0.453</td>
</tr>
</tbody>
</table>
Which features to use?

All features
- Word context
  - Starts uppercase
  - Contains space
  - Is number
  - Contains letters
  - All capital letters
  - Single letter
  - Contains punctuation
  - Part of speech
- Document context
  - Tf
  - First offset
  - First sentence offset
  - Sentence importance (SumBasic)
  - Dispersion (Gries, 2008)
- Collection context
  - Idf
  - Tf.idf
  - Is training keyword
  - Assignment ratio

Minimum set
- Part of speech
- Dispersion
- Tf.idf
- Assignment ratio

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>P@5</th>
<th>P@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank-SVM</td>
<td>0.399</td>
<td>0.631</td>
<td>0.453</td>
</tr>
<tr>
<td>minimum set</td>
<td>0.405</td>
<td>0.631</td>
<td>0.459</td>
</tr>
</tbody>
</table>
Conclusion

• For the Dutch Folktale Database
  • Uncontrolled indexing is necessary
  • Many single word keywords which appear (almost) literally in text
  • Moderate to substantial agreement between annotators
• Learning to rank can be used for suggesting keywords
  • 3 out of top 5 relevant
  • Important features:
    1) assignment ratio, 2) tf.idf, 3) part-of-speech and 4) dispersion
• Future work
  • Deal with multilingual content
  • Suggest abstract keywords
Questions?

• D.Trieschnigg@utwente.nl