A quantitative evaluation of keyword measures for corpus-based discourse analysis

Stefan Evert, Natalie Dykes, Joachim Peters
FAU Erlangen-Nürnberg

www.linguistik.fau.de
Aim

- Keywords as “a quick and simple ‘way in’” to corpus comparison (Baker et al. 2013)
- Previous approaches to KW calculation focus on mathematical adequacy and/or number of generated items (cf. Kilgarriff 2001, Paquot & Bestgen 2009, Lijffijt et al. 2016)

Our approach:
- Previously determined qualitative linguistic categories
- Evaluate statistically generated keyword lists against them
- Procedure specifically tailored to discourse analysis
Corpus

- 14.3M token corpus on German web data about multi-resistant pathogens (MRO) collected with BootCat (Baroni & Bernardini 2004)
- 9,750 texts of varying genres and lengths
- Overall corpus metadata (manual)
  - Actor: author
  - Actor: intended reader
  - Topic
    - MRO
    - related topic (clinical hygiene; other infections...
Corpus

Extraction of relevant subcorpus via metadata

- Actor – author: media
- Actor – reader: general public
- Topic: MRO

1,3M tokens (1,177 texts) of mass media texts and reader comments taken from the MRO corpus
Reference corpora

- Years 2011–2014 of *Süddeutsche Zeitung* (SZ), a left-leaning daily newspaper (290M tokens)
- Years 2011–2014 of *Frankfurter Allgemeine Zeitung* (FAZ), a right-leaning daily newspaper (150M tokens)
- All corpora: POS-tagged with TreeTagger and lemmatised with SMOR (Schmid et al. 2004)
Annotation categories

Annotation of top 200 lexical KW for different techniques following gold standard based on previous analysis of a different MRO press corpus (Peters 2017)

Adaption of selected aspects of the DIMEAN model (Spitzmüller/Warnke 2011)

- Actor
- Topos
- Metaphor
- False positives (unclear/other/irrelevant)
- Additional category: evaluative lexis (positive/negative stance)
Annotation procedure

Web-based annotation platform MiniMarker
Agreement

- Two independent annotators
- Agreement of 82.2% on distinction TP vs. FP (but Cohen $\kappa = .566$ fairly low)
- Domain-specific, highly frequent words often marked FP (“unclear”) by one annotator and TP by the other
- Disagreements between TP categories less frequent; mostly due to overlap between discourse levels
  - metaphors as part of topoi
  - intertwined argumentational levels
- Final gold standard jointly reconciled by annotators
### Keyword extraction techniques

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1 - n_1$</td>
<td>$f_2 - n_2$</td>
<td></td>
</tr>
</tbody>
</table>

- $f_1 = \text{freq. in target corpus}$
- $n_1 = \text{sample size of target}$
- $f_2 = \text{freq. in reference corpus}$
- $n_2 = \text{sample size of reference}$

- **Textbook approach:** $G^2$ log-likelihood significance test (Dunning 1993)
- **Effect size measure:** $LR$ log ratio $f_1/n_1 : f_2/n_2$ (Hardie unpublished)
  - combined with Bonferroni-corrected significance filter
- **Statistician's choice:** $LR_{\text{cons}}$ conservative LR (Evert p.c.)
  - lower bound of confidence interval (Hardie's formula)
  - with Bonferroni correction
Keyword extraction techniques

<table>
<thead>
<tr>
<th>$f_1$</th>
<th>$f_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1 - n_1$</td>
<td>$f_2 - n_2$</td>
</tr>
</tbody>
</table>

- $f_1 = \text{df in target corpus}$
- $n_1 = \#\text{texts in target corpus}$
- $f_2 = \text{df in reference corpus}$
- $n_2 = \#\text{texts in reference}$

- Methodological discussion: non-randomness / term clustering as key issue
- Simple correction: use document frequency (df) instead of raw frequency
- Mathematical justification as statistical inference for $\alpha$ parameter of Katz (1996)
Experiments

- Extract top-200 keywords for each technique
  - frequency threshold $f \geq 5$ in reference corpus, because we are not interested in terminology extraction
- Manual annotation of TPs (categories, evaluative)
- Two comparable reference corpora: Süddeutsche (SZ) vs. Frankfurter Allgemeine (FAZ)
- Keywords based on raw frequency (classic) vs. document frequency (df-based)
Overlap between techniques

Süddeutsche Zeitung (SZ)

G2: 101
LRcons: 62
LR: 60

Frankfurter Allgemeine (FAZ)

G2: 103
LRcons: 66
LR: 57
Overlap between techniques

**Extraction techniques**

- SZ
- FAZ
- SZ (df)
- FAZ (df)

**SZ vs. FAZ**

- Classic
- DF-based

**Classic vs. DF-based**

- SZ
- FAZ

---

LR / LRcons  LR / G2  LRcons / G2  G2  LRcons  LR  G2  LRcons  LR
Frequency bias
Precision = \#TP / 200 cand.

TP = assigned to category and/or evaluative
Recall = \#kw for each category

Keywords against FAZ (classic)
Recall = \#kw for each category
Recall = #kw for each category
Why so few metaphor keywords?

Possible causes:
- No metaphors in online media discourse (unlikely)
- Cannot be reduced to single words
- Keywords occur, but are too infrequent
A case study

- List of plausible keywords for each metaphor category from thesaurus (Dornseiff 2004)
  - e.g. POLICE: *Indiz* clue, *Killer* killer, *Mord* murder, *Täter* culprit, *fahnden* search, *heimtückisch* insidious, ...
  - manually validated against concordance in target corpus

- Comparison with full set of keyword candidates
  - frequency in target corpus
  - removed because of reference corpus threshold?
  - keyness score and rank in candidate set
A case study

Dornseiff metaphor keywords in MRSA corpus

- **metaph: water**
- **metaph: war**
- **metaph: tech**
- **metaph: space**
- **metaph: police**
- **metaph: game**
- **metaph: economy**

Number of keywords:
- Dark grey: \( f \geq 10 \)
- Light grey: \( f \geq 5 \)
- White: \( f < 5 \)
Finding metaphor keywords

- Substantial number of plausible keywords for all metaphor categories except **ECONOMY**
  - frequent in target corpus & pass threshold in reference
  - but very low ranks (> 1000) from all keyness measures
- Reason: literal senses very frequent in reference
  - aggregating all keywords from category doesn't help
- Approximate semantics with distributional context vectors (Schütze 1998)
  - three-sentence context around each potential keyword
  - bag-of-words centroids of word embeddings
  - MRSA contexts clearly separated from reference contexts?
Finding metaphor keywords

Kampf

angreifen
Finding metaphor keywords
Conclusion

- Quantitative evaluation of keyword techniques & parameters for corpus-based discourse analysis
- Small overlap between $G^2$ and LR keywords – but choice of reference corpus makes little difference
- All techniques achieve high precision $> 80$
- Recommendation: $LR_{\text{cons}}$ on document frequency
- Good recall for some categories, poor for metaphors
- Suitable keywords *are* available $\rightarrow$ new techniques

And Thank You for your attention!

Don't forget Natalie & Joachim's talk at 14:20 (same room).
References


References


Annotation scheme

Categories from previous manual study on smaller corpus (Peters 2017)

**Metaphors**
- machines
- war
- control
- police/crime
- games/sports
- space
- water
- economy

**Actors**
- patients
- medical staff
- scientists
- politicians
- pharmaeutics
- hospital
- pathogen
Annotation scheme

Topics

General
- Medical history
- Evolution
- Other countries
- Spread

Causes for MRO
- Working conditions
- Economic efficiency
- Pharmaceutical companies
- Treatment
- Negligence
- Unethical actions
- Livestock

Solutions
- Hospitals
- Economy
- Structural changes

Media
- Self-reflection
- Disaster scenarios
Confusion matrix (primary category)