

# Evaluating the Reliability and Interaction of Recursively Used Feature Classes for Terminology Extraction

Anna Hättü<sup>1,2</sup>

Michael Dorna<sup>1</sup>

Sabine Schulte im Walde<sup>2</sup>

<sup>1</sup>Corporate Research, Robert Bosch GmbH

<sup>2</sup>Institute for Natural Language Processing (IMS), University of Stuttgart

## Automatic Term Recognition

**terms:** linguistic units which characterize a specific topic domain  
 → e.g. term candidates in computational linguistics domain: *parsing, machine translation* and *natural language generation*

## Approach

- **goal:** improve automatic term recognition
- **common approach:** use a statistical measure to find terms
- **this work:** take several statistical measures as features to a classifier (→ Random Forest and Decision Tree)

## Data

- ACL RD-TEC 1.0
- benchmark dataset for term extraction
- corpus of 10,922 ACL publications
- manual annotation of 22,044 valid terms and 61,758 non-terms

## Automatic Term Recognition Pipeline

extract uni-, bi- and trigrams from text

compute features

random forest classifier

feature reduction

decision tree classifier

10-fold cross validation

## Features

### Term-Document Measures

- variants of tf-idf
- term variance ...

### Domain-Specificity Measures

- weirdness ratio
- TFITF...

### Association Measures

- local mutual information
- chi square

### Count-based Measures

- frequency
- word length

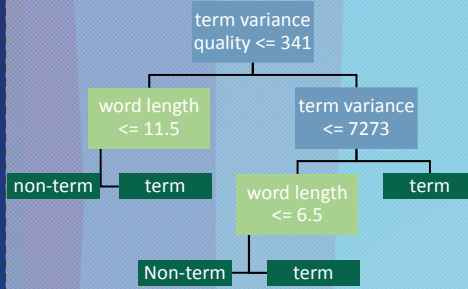
### Linguistic Feature

- POS-tag pattern

### Features of Components

- → features above are applied to substrings:
- 0-uni-CSmw
- 1-bi-CSmw, ...

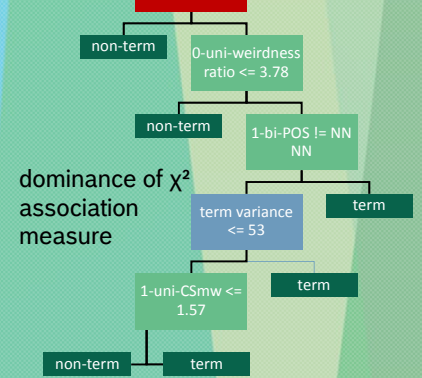
## Unigrams



combination of term-doc and count-based measures

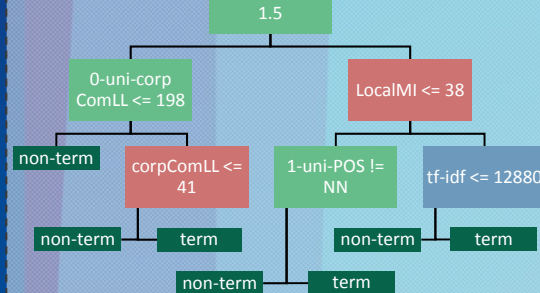
special cases: *czech, newspaper, chain, ...*

## Trigrams



dominance of  $\chi^2$  association measure

## Bigrams

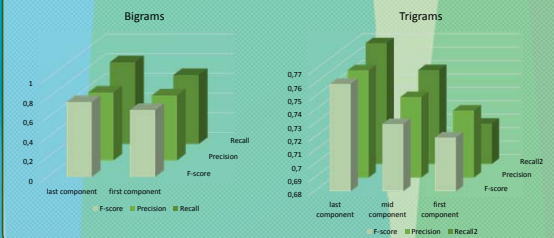


combination of various classes

why component feature as head node?  
 → *this drop, general scheme, particular cluster, ...*

## Component Class

- only "valid" and "non-valid" term candidates
- only candidates where both classes have same components
- → *bayes algorithm* (valid) vs. *basic algorithm* (non-valid)



## Results

Method	F1	Method	F1	Method	F1
<b>best single feature</b> (term variance qual.)	0.70	<b>best single feature</b> (0-uni-CSmw)	0.72	<b>best single feature</b> (Chi <sup>2</sup> )	0.90
<b>best feature class</b> (term-doc metrics)	0.72	<b>best feature class</b> (component features)	0.80	<b>best feature class</b> (component features)	0.92
<b>all features</b>	<b>0.77</b>	<b>all features</b>	<b>0.81</b>	<b>all features</b>	<b>0.93</b>

Unigrams

Bigrams

Trigrams

## Conclusion

- a combination of measures improves the system
- single-word terms are harder to recognize than multiword terms
- for terms of different length, different measure combinations should be used
- for determining termhood, modifiers are better indicators than heads