LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Developing a stemmer for German based on a comparative analysis of publicly available stemmers

Leonie Weißweiler, Alexander Fraser
CIS, LMU Munich
GSCL 2017, September 13th

Outline

- Introduction to Stemming
- Overview of existing stemmers
- Objectives
- Gold standard development and evaluation methodology
- Evaluation results
- CISTEM development
- Final evaluation
- Conclusion

The Stemming Task

"In Information Retrieval, an important task is to not only return documents that contain the exact query string, but also ***documents containing semantically related words or different morphological forms*** *of the original query word"*

(Manning et al., 2008, p.57)

"A stemming algorithm is a computational procedure which ***reduces all words with the same root*** *(or, if prefixes are left untouched, the same stem)* ***to a common form****, usually by stripping each word of its derivational and inflectional suffixes"*

(Lovins, 1968)

Motivation

- Small choice of stemmers for German available
- Snowball is the most common in NLP toolkits
- No evaluation of different stemmer performances available

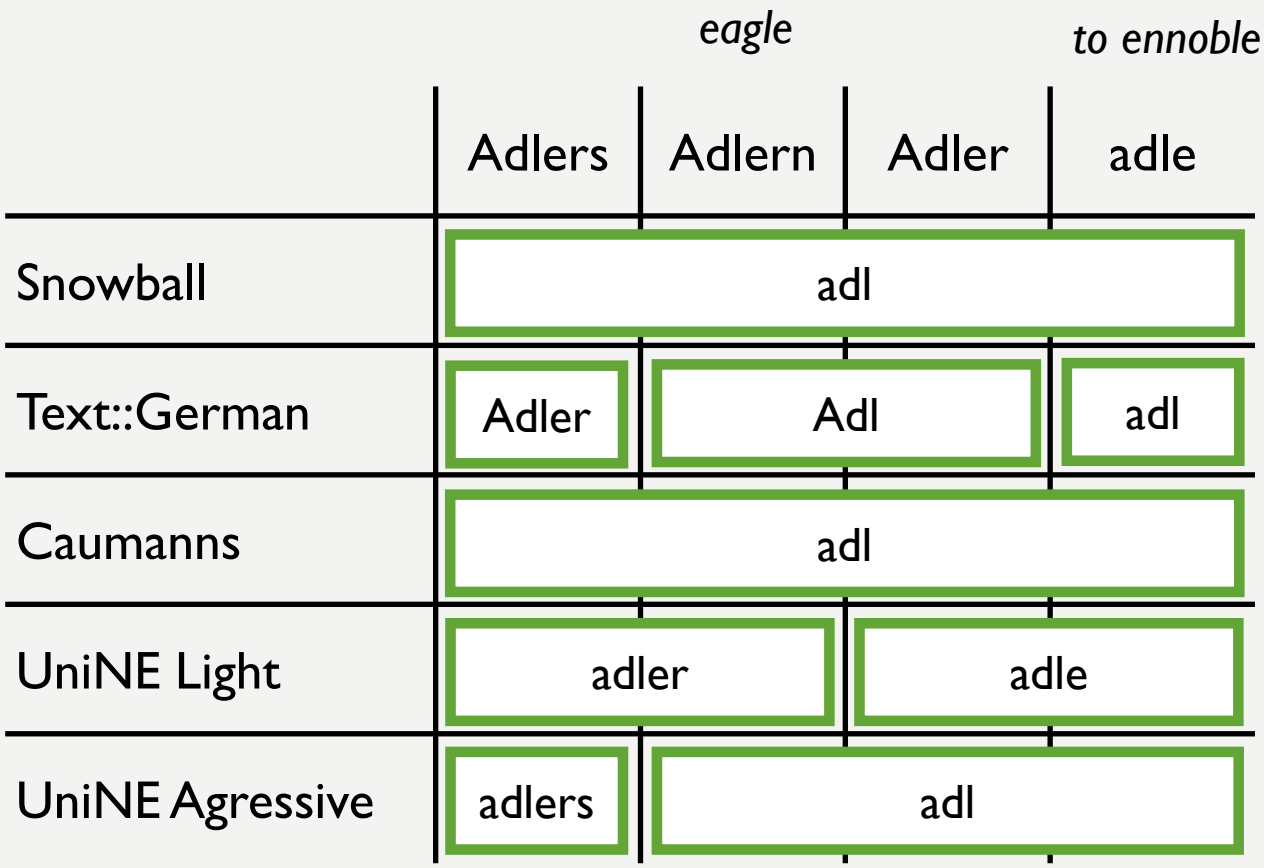
Existing Stemmers

- Snowball
- Text::German
- Caumanns
- UniNe (Light or Aggressive)

Objectives

- Present first comparative evaluation of existing stemmers for German
- Present new state-of-the-art stemmer based on evaluation results
- Make official implementations available in a range of programming languages

Comparison


eagle
to ennoble
Adlers
Adlern
Adler
adle
Snowball
adl
Text::German
Adler
Adl
adl
Caumanns
adl
UniNE Light
adler
adle
UniNE Agressive
adlers
adl

Gold standard 1

Word	Belichtungsmesser					
Morphemes	be	licht	ung	s	mess	er
Shapes	x	A	x	x	V	x

light meter

Gold standard 1

Word	Belichtungsmesser					
Morphemes	be	licht	ung	s	mess	er
Shapes	x	A	x	x	V	x

light meter

Gold standard 1

Word	Belichtungsmesser
Stem	lichtmess

light meter

Gold standard 1

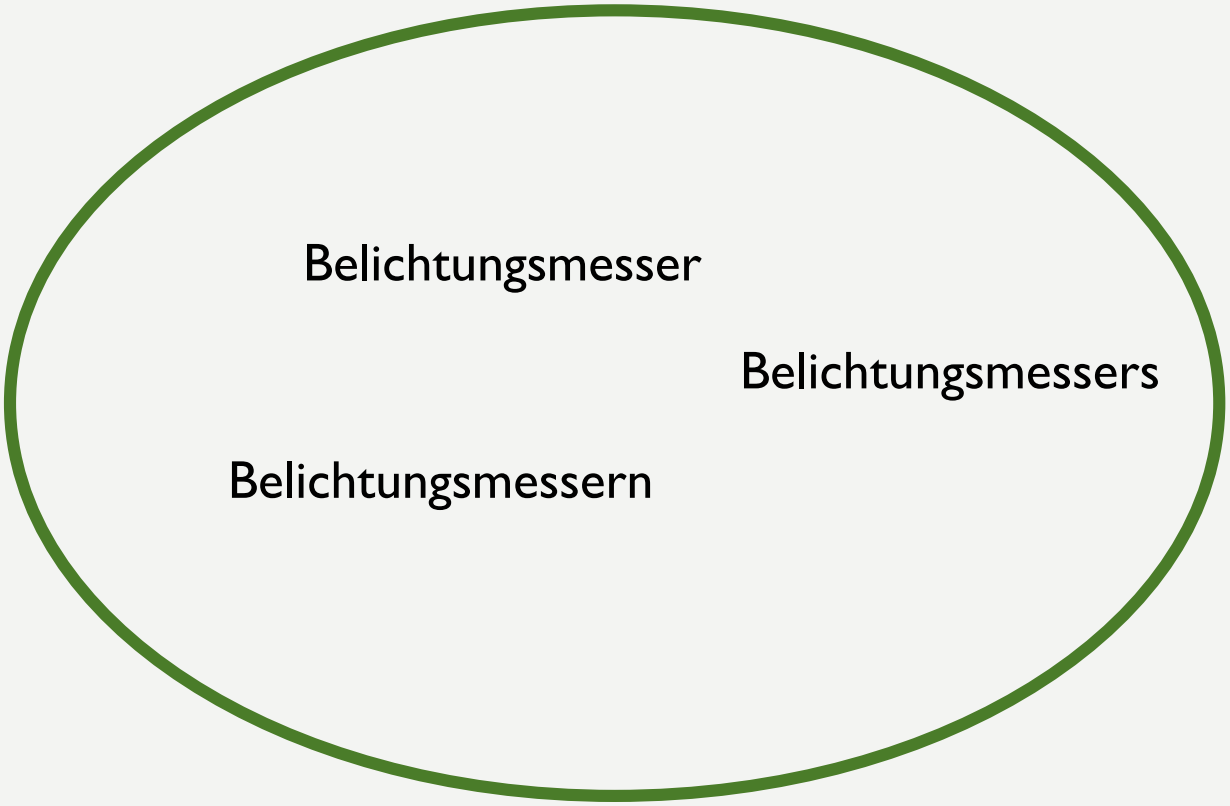
Word	Belichtungsmesser
Stem	lichtmess

Word	Belichtungsmessers
Stem	lichtmess

Word	Belichtungsmessern
Stem	lichtmess

light meter

Gold standard I

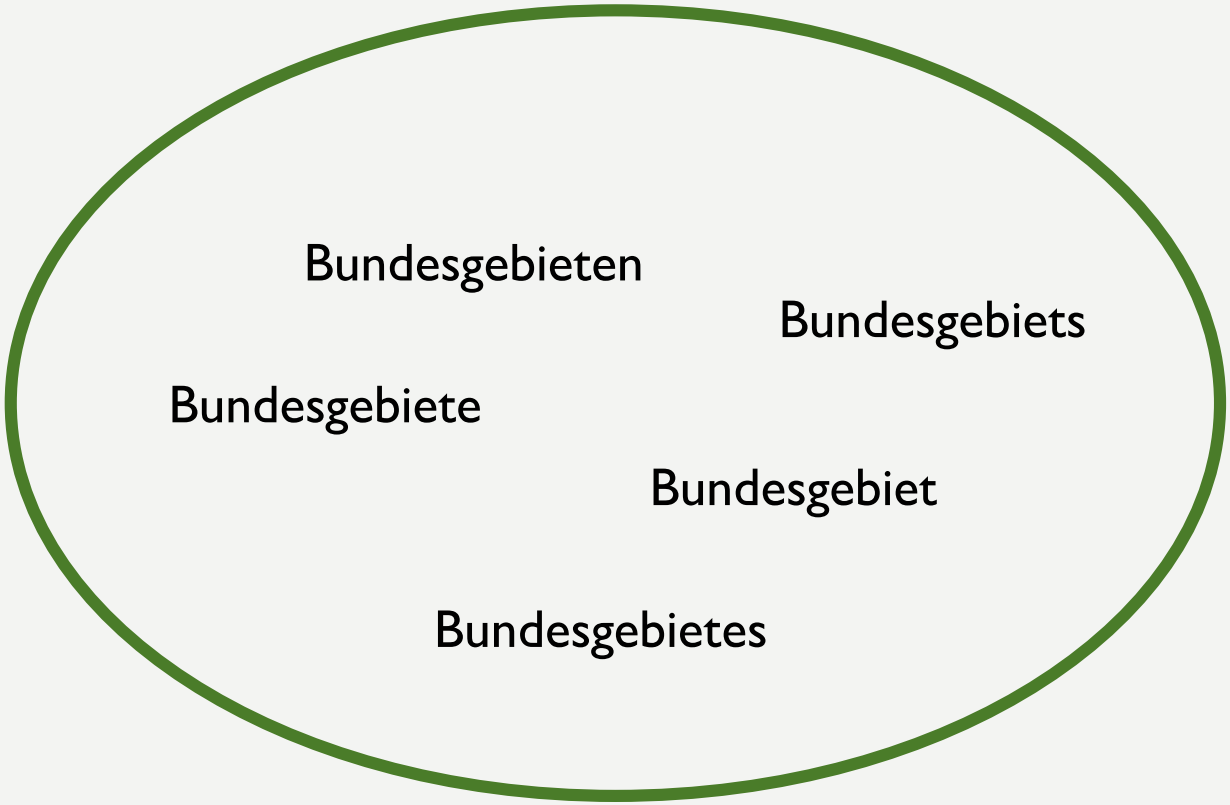


light meter

Gold standard 2

- Uses wordform – lemma information from the CELEX2 corpus
- Groups wordforms with the same lemma
- More conservative that gold standard 2

Gold standard 2



federal territory

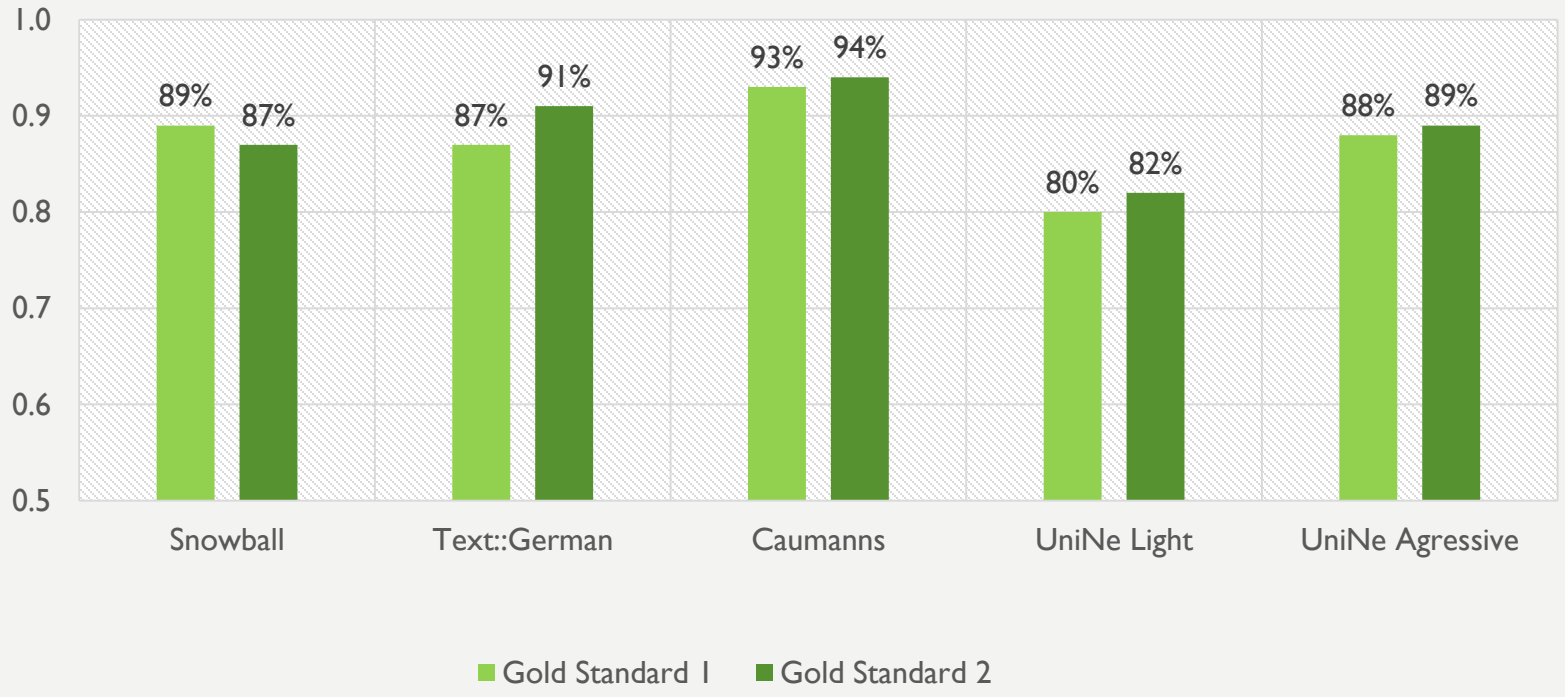
Goldstandard 1 vs 2

Goldstandard 1		Goldstandard 2
relativem Relatives … …	relative	relativieret releativiertest …
Relativität Relativitäten … …	the theory of relativity	Relativität Relativitäten
relativistischerer …	an idea in philosophy	Relativismus relativistischen …

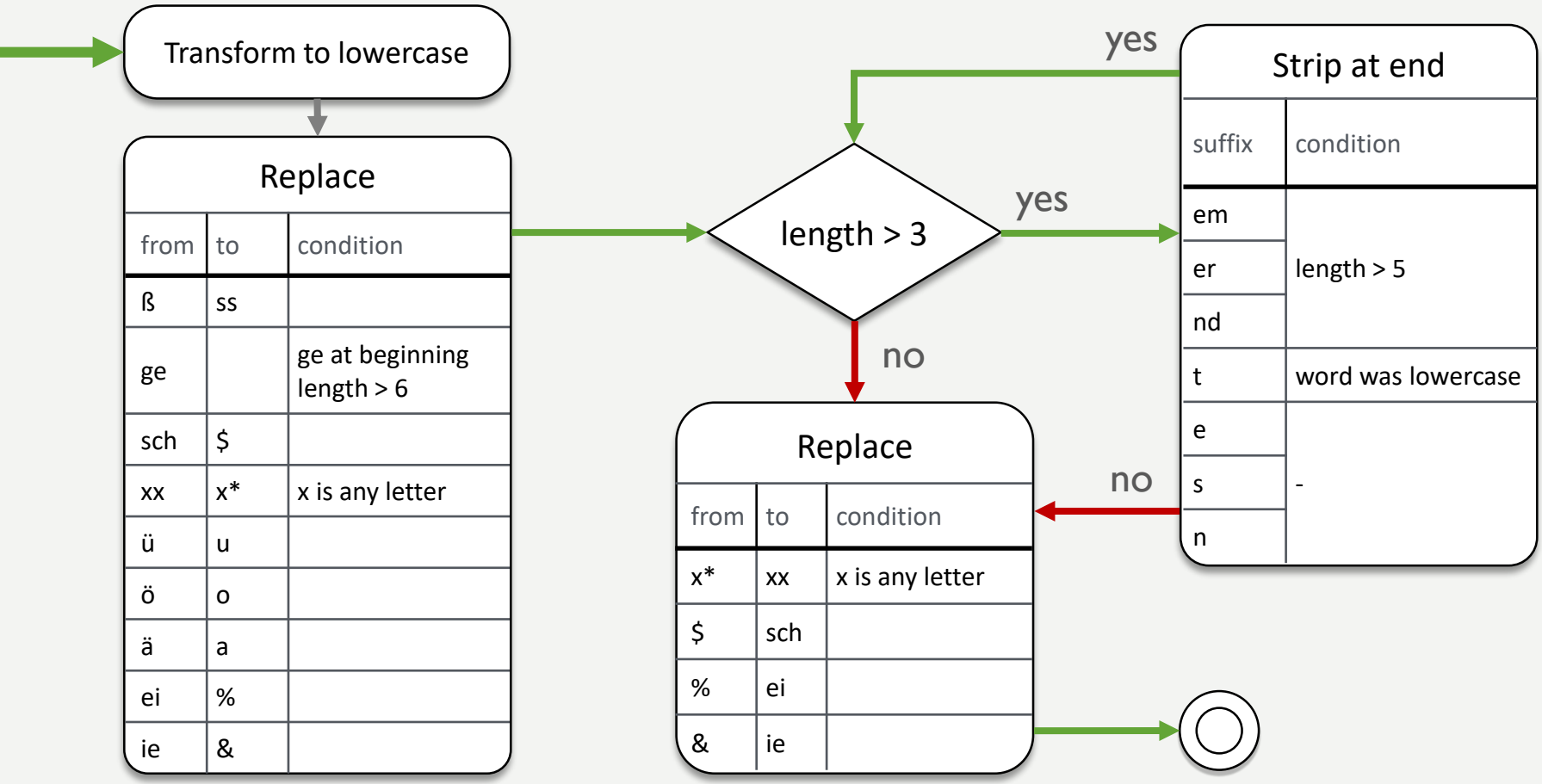
Evaluation Methods

- Stem the CELEX2 Corpus
- For each stem/cluster/line in each gold standard, compute precision and recall by finding the stem in the stemmed corpus that best matches it
- Compute average precision and recall for each gold standard + f1 measure

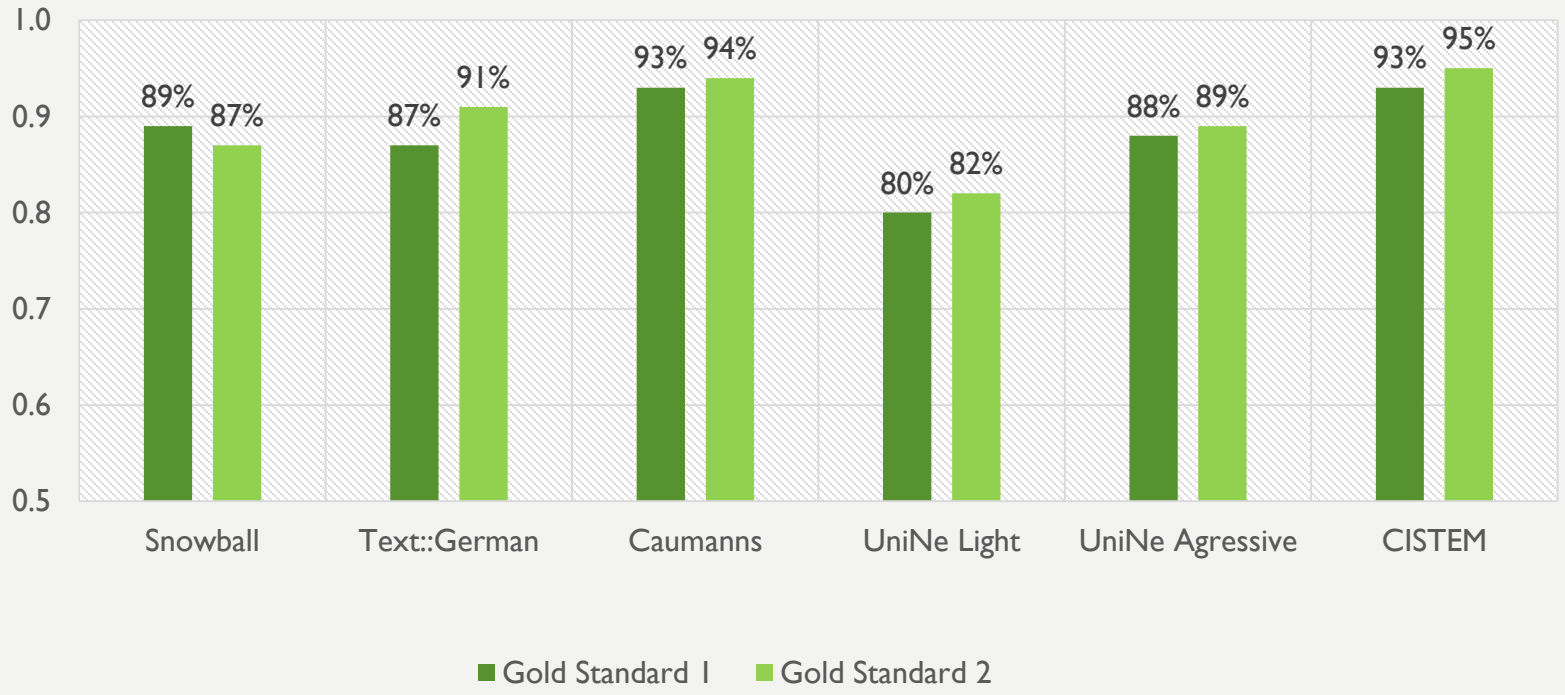
F1 Scores


1.0
0.9
0.8
0.7
0.6
0.5
89%
87%
87%
91%
93%
94%
80%
82%
88%
89%
Snowball
Text::German
Caumanns
UniNe Light
UniNe Agressive
Gold Standard 1
Gold Standard 2

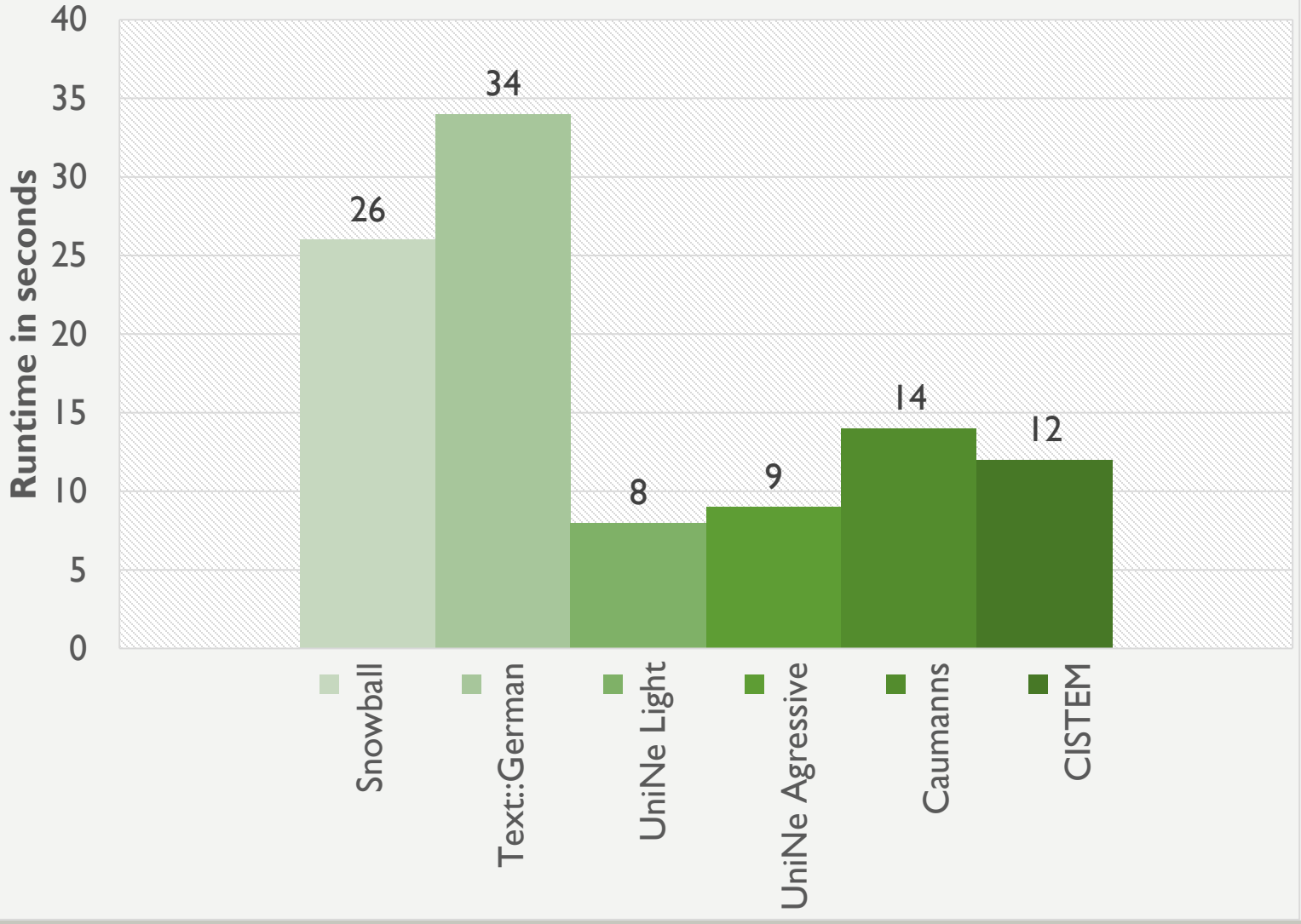
CISTEM Development



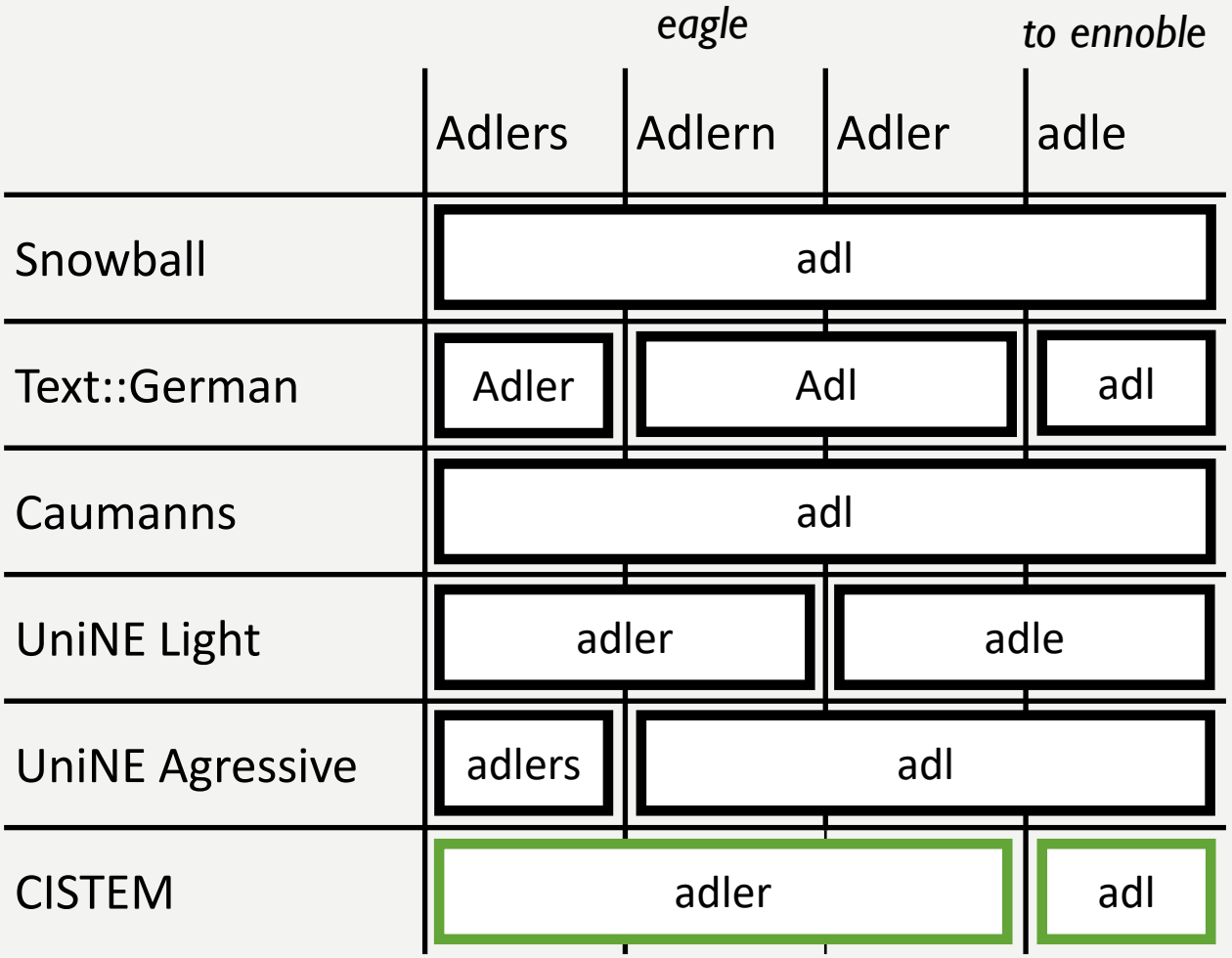
F1 Scores


1.0
0.9
0.8
0.7
0.6
0.5
89%
87%
87%
91%
93%
94%
80%
82%
88%
89%
93%
95%
Snowball
Text::German
Caumanns
UniNe Light
UniNe Agressive
CISTEM
Gold Standard 1
Gold Standard 2

Runtime


Runtime in seconds
0
5
10
15
20
25
30
35
40
26
34
8
9
14
12
Snowball
Text::German
UniNe Light
UniNe Agressive
Caumanns
CISTEM

Comparison


eagle
to ennoble
Adlers
Adlern
Adler
adle
Snowball
adl
Text::German
Adler
Adl
adl
Caumanns
adl
UniNE Light
adler
adle
UniNE Agressive
adlers
adl
CISTEM
adler
adl

Conclusion

CISTEM…

- ... scores highest in F1 measure
- ... is one of the quickest stemmers (when implemented in Perl)
- ... is easy to use and understand
- ... has a case insensitive version
- ... has official implementations available in Python, Java, Perl and C with more to come

Future Work

- Release official CISTEM implementations in as many programming languages as possible
- Try to merge the two gold standards into one definitive gold standard
- Implement rule-learning system for learning the optimal stemmer (would require definitive gold standard)
- Test if testing the gold standards yield the same results as testing stemmers in an IR system

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thanks for your attention!
github.com/LeonieWeissweiler/CISTEM

This project has received funding from the European Unions Horizon 2020research and innovation programme under grant agreement No 644402 (HimL).This projecthas received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 640550).

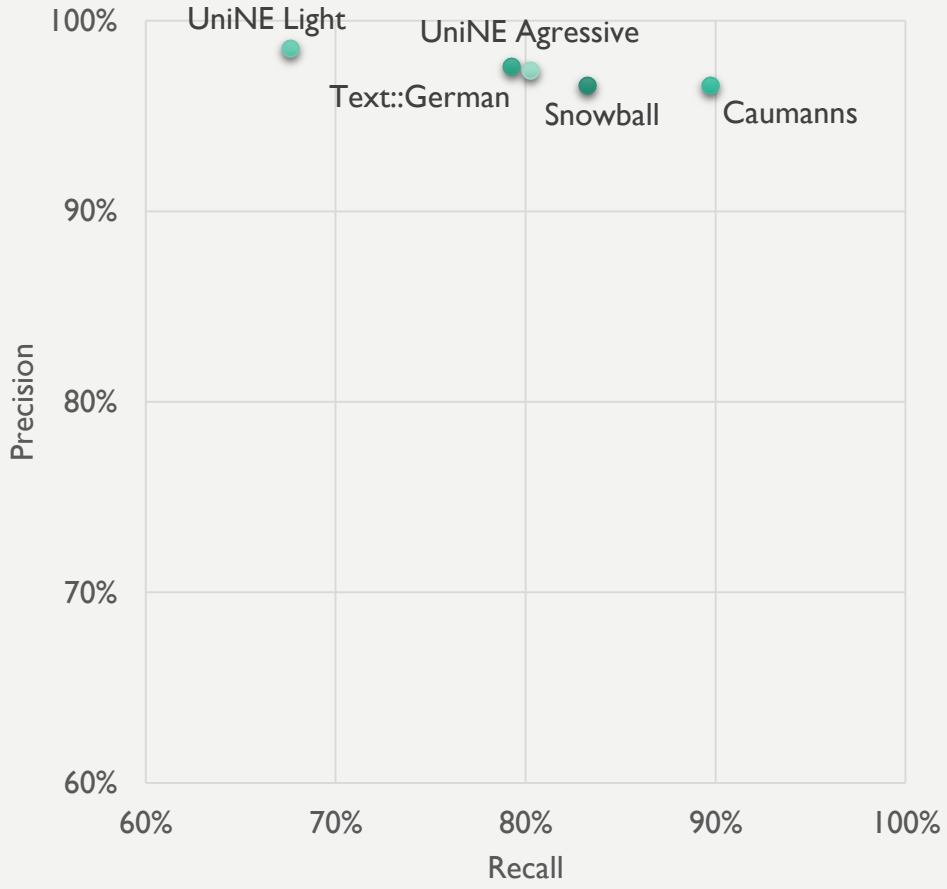
erc

Health in my Language

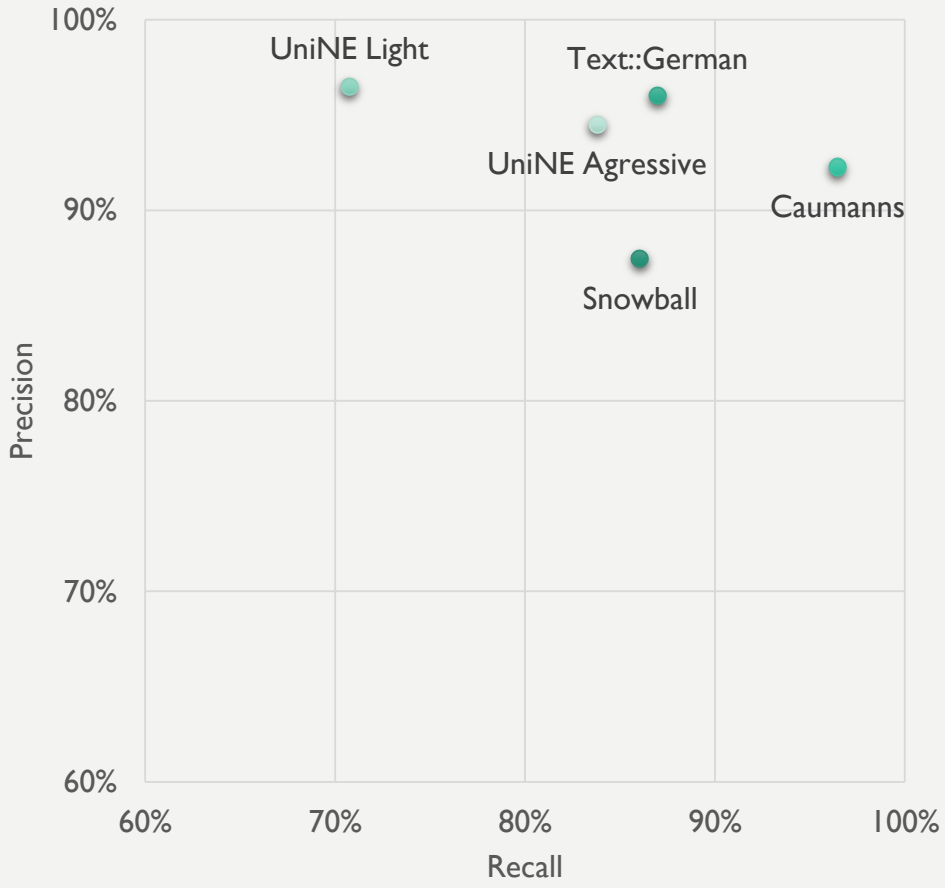
HimL

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

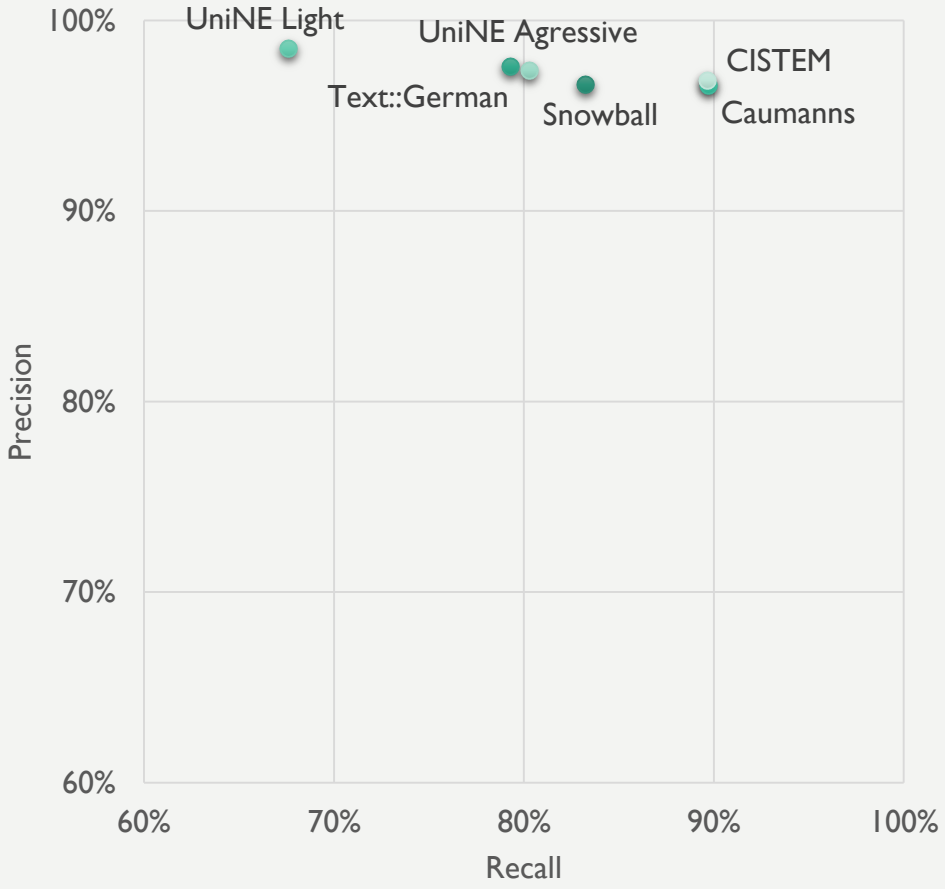
Precision & Recall Gold standard I



Precision & Recall Gold standard 2



Precision & Recall Gold Standard I



Precision & Recall Gold Standard 2