SoMaJo: State-of-the-art tokenization for German web and social media texts

Thomas Proisl [1] Peter Uhrig [2]
10th Web as Corpus Workshop (WAC-X)
[1]Professur für Korpuslinguistik
[2]Lehrstuhl für Anglistik, insbesondere Linguistik

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Tokenization – a simple and boring task?

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

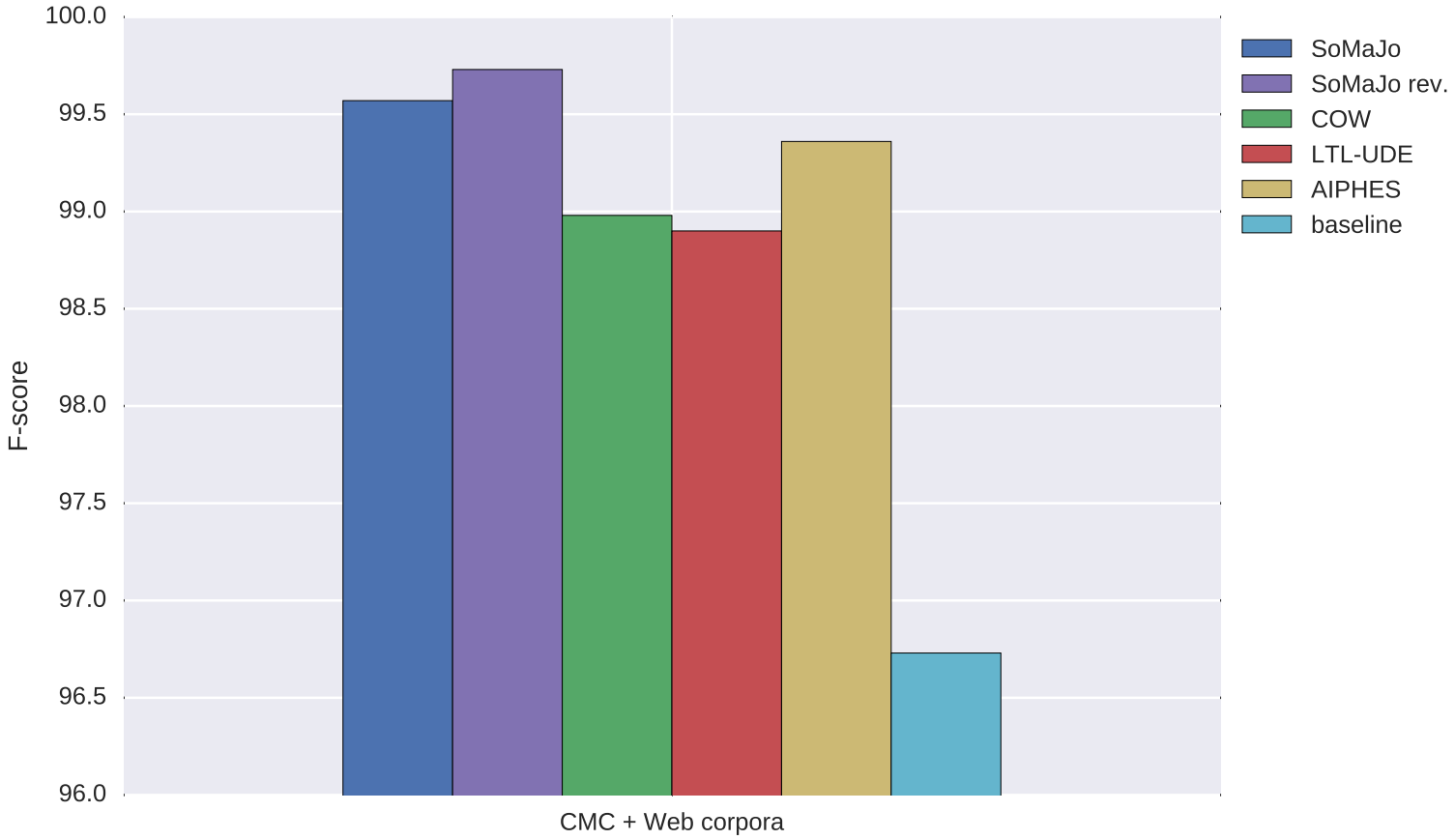
Ad-hoc approach to tokenization

- tokenization is usually the first processing step in layered NLP pipelines
- foundation for all later processing steps
- often seen as a boring and trivial task
- unproblematic for humans
  - good at pattern finding
  - can work with ambiguity
- common ad-hoc approach: use simple regular expressions
- our baseline is a simple sed one-liner:

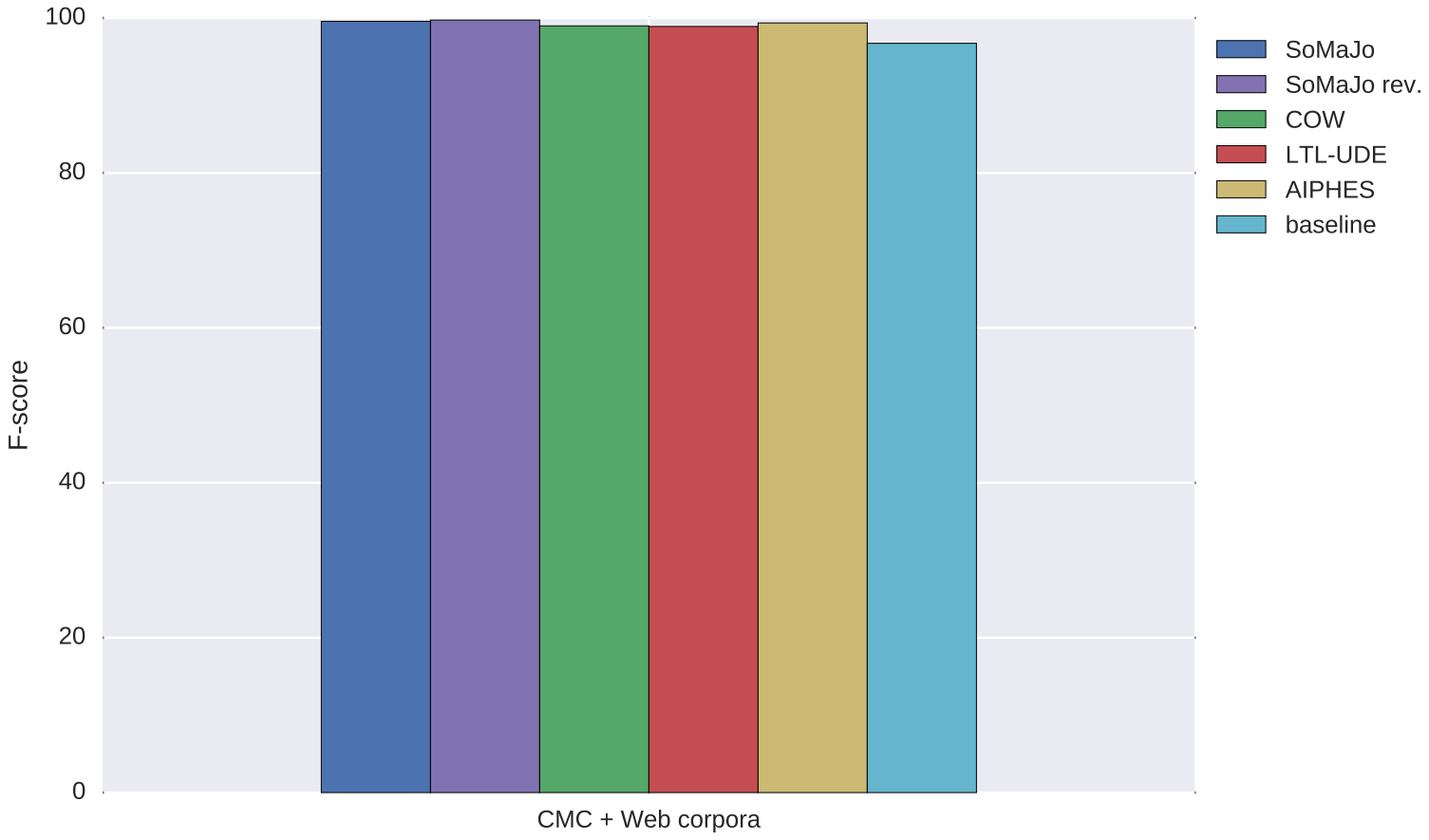
```
sed -re "/^<[^>]+>$/! { s/([.!?,;:+*()\"'-])/ \1 /g; s/\s+/\n/g }"
```

- results look satisfactory: average F-score of 96.73 (94.91 CMC, 98.55 web corpora)

Dramatic differences between the systems…



… or not


100
80
60
40
20
0
F-score
CMC + Web corpora
SoMaJo
SoMaJo rev.
COW
LTL-UDE
AIPHES
baseline

Tokenization guidelines

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

General overview

- usual treatment of whitespace and punctuation
  - proper names are not split up: `H&M` → `H&M`
- abbreviations:
  - multidot abbreviations representing multiple words are split up: `d.h.` → `d. | h.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`
- single tokens:
  - HTML/XML tags
  - e-mail addresses
  - URLs
  - filenames
  - emoticons

Controversial cases: Contractions and whitespace

- contracted forms: `mit'm Fahrrad` → `mit'm | Fahrrad`
  - compare conventional English segmentation of `I'm` → `I | 'm` or `don't` → `do | n't`
  - more difficult to find individual words; multiple tags and lemmata have to be assigned to one token or introduction of new tags
- whitespace errors:
  - not usually corrected: `schona ber` → `schona | ber`
  - except in emoticons and punctuation:
    - `tag quaki : )` → `tag | quaki | :)`
    - `f - > d` → `f | -> | d`
  - what about `d < - f`?

Controversial cases: Dates

- `1980-07-21` → `1980 | -07 | -21`
  - strange token `-07`
  - SoMaJo can tell you that all three tokens are part of a date:

```
echo "1980-07-21" | python bin/tokenizer -t -
1980    date
-07     date
-21     date
```

Controversial cases: CamelCase

- CamelCase: omitted space vs. proper name
  - `Zu welchemHandlungsbereich gehört` → `... | welchem | Handlungsbereich | ...`
  - `InterCity` → `InterCity`
  - CamelCase-splitting is optional in SoMaJo!
    - Wiki syntax
    - naming conventions in programming languages such as Java, C#
    - study of non-conventional spellings

Ambiguous cases

- horizontal ellipsis:
  - `du bist echt ein Arm...` → `Arm...`
  - `zeig mir mal deinen Arm...` → `Arm | ...`
- fraction vs. division:
  - `1/2 Tasse` → `1/2`
  - `1/2+5=x` → `1 | / | 2`
- typographical errors:
  - `Laub- und Nadelbäume` → `Laub- | und | Nadelbäume`
  - `handfest un direkt- so sind se` → `... | direkt | - | ...`

System description

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

General approach

- rule-based: cascade of regular expressions
- additional lexical resources
- replacement of “problematic” tokens with unique pseudotokens to prevent subsequent rules from processing them any further
- can output the token class for each token, e.g. number, XML tag, abbreviation, etc.

→ For details, visit our poster!

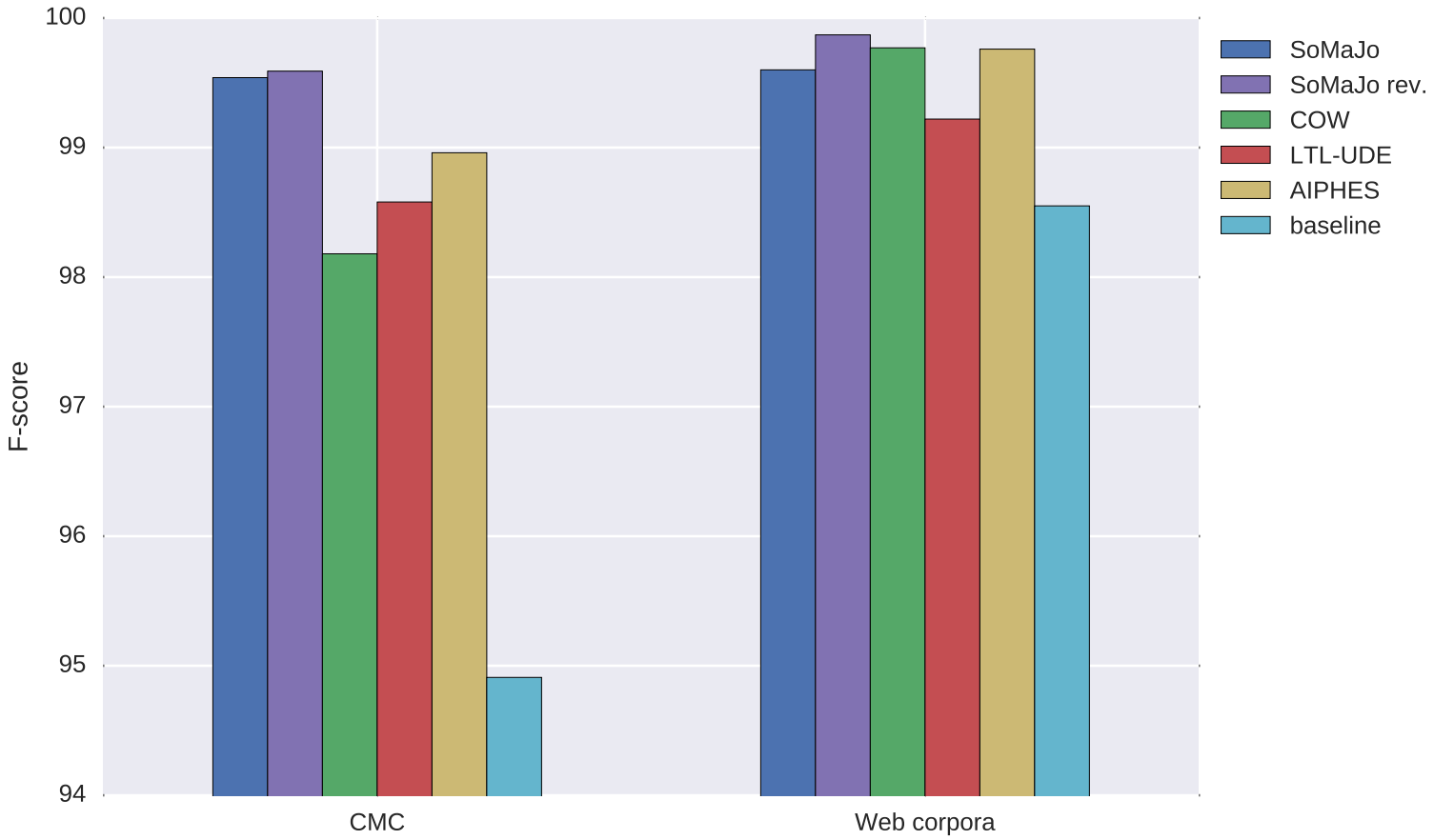
Results

Results

	CMC			Web corpora			macro average
	P	R	F	P	R	F	F
baseline	91.84	98.20	94.91	98.27	98.84	98.55	96.73
submission	99.52	99.56	99.54	99.57	99.64	99.60	99.57
revised	99.62	99.56	99.59	99.83	99.92	99.87	99.73

- SoMaJo outperforms all other systems in the overall shared task
- ranks first on the CMC dataset and third on the web corpora dataset
- revised version would also rank first on the web corpora dataset

Results


100
99
98
97
96
95
94
F-score
CMC
Web corpora
SoMaJo
SoMaJo rev.
COW
LTL-UDE
AIPHES
baseline

Error analysis

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Remaining problems: Inherently ambiguous cases

- hyphens (-):
  - typographically correct use for example in *Bindestrichkomposita*: `Harry-Potter-Roman` → `Harry-Potter-Roman`
  - some people use it as a *Bis-Strich* instead of the typographically correct en dash (–): `von elf-zwölf` → `von | elf | - | zwölf`
- abbreviations vs. actual words: `automat.`, `zum.`
- cardinal number at end of sentence vs. ordinal number
- section numbers vs. dates: `5.3.`
- citations (`Storrer2007`) vs. proper names (`Blume2000`)

Remaining problems: Rare and unsystematic problems

- omitted whitespace:
  - `bei WAHLTEHMEN.DEteilzunehmen` → `bei | WAHLTHEMEN.DE | teilzunehmen`
- usernames: `Erdbeere$`
- abbreviations without dots: `zB`, `idR`
- multiple stars: `angewandt?**` → `angewandt | ? | **`
- proper names: `tacheles.02spezial`
- double brackets: `[[security:verschlüsselung]]` → `[[ | security | : | verschlüsselung | ]]`

Conclusion

Conclusion

- tokenization as such is a relatively simple task
  - all teams scored very high
  - even a trivial baseline performs over 96 F-score.
- making any further improvements is an extremely laborious task
  - Zipfian distribution of the items is to be expected
- SoMaJo is freely available: `https://pypi.python.org/pypi/SoMaJo`

Questions!

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE