Inflection-Tolerant Ontology-Based Named Entity Recognition for Real-Time Applications

Christian Jilek Markus Schröder Rudolf Novik

Sven Schwarz Heiko Maus Andreas Dengel

Parts of this work were funded by the DFG in the SPP on Intentional Forgetting in Organizations.

Motivation

2013 – 2016 (EU project)

2016 – 2019 (DFG project)

Forgetful & Self-Organizing Information Systems

(to support information management & knowledge work)

↓

continuous **information value assessment**

↓

continous **user activity tracking** and **evidence processing**

↓

information extraction in (near) real-time

Problem of Inflections

Aussagenlogik

Die **Aussagenlogik** ist ein Teilgebiet der Logik, das sich mit Aussagen und deren Verknüpfung durch Junktoren befasst, ausgehend von strukturlosen Elementaraussagen (Atomen), denen ein Wahrheitswert zugeordnet wird. In der klassischen Aussagenlogik wird jeder Aussage genau einer der zwei Wahrheitswerte „wahr" und „falsch" zugeordnet. Der Wahrheitswert einer zusammengesetzten Aussage lässt sich ohne zusätzliche Informationen aus den Wahrheitswerten ihrer Teilaussagen bestimmen.

Confidence: 0.5 Language: German

Die Aussagenlogik ist ein Teilgebiet der Logik, das sich mit Aussagen und deren Verknüpfung durch Junktoren befasst, ausgehend von strukturlosen Elementaraussagen (Atomen), denen ein Wahrheitswert zugeordnet wird. In der klassischen Aussagenlogik wird jeder Aussage genau einer der zwei Wahrheitswerte „wahr" und „falsch" zugeordnet. Der Wahrheitswert einer zusammengesetzten Aussage lässt sich ohne zusätzliche Informationen aus den Wahrheitswerten ihrer Teilaussagen bestimmen.

Confidence: 0 Language: German

Die Aussagenlogik ist ein Teilgebiet der Logik, das sich mit Aussagen und deren Verknüpfung durch Junktoren befasst, ausgehend von strukturlosen Elementaraussagen (Atomen), denen ein Wahrheitswert zugeordnet wird. In der klassischen Aussagenlogik wird jeder Aussage genau einer der zwei Wahrheitswerte „wahr" und „falsch" zugeordnet. Der Wahrheitswert einer zusammengesetzten Aussage lässt sich ohne zusätzliche Informationen aus den Wahrheitswerten ihrer Teilaussagen bestimmen.

DBpedia Spotlight:

P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In Proc. of the 7th Int'l Conf. on Semantic Systems (I-Semantics), pages 1–8. ACM, 2011.

Related Work

Inflection-Tolerant NER	Real-Time Capable NER
Savary & Piskorski (2010) → IE platform SProUT, Polish, explicitly listing all inflected forms	**Dlugolinsky, Nguyen et al. (2013/2014)** → several gazetteer-based approaches
Day & Prukayastha (2013) → NER for Indian languages, gazetteer-based & ML & hybrid	**Al-Rfou & Skiena (2012)** → SpeedRead, 10x faster than CoreNLP, 153 tokens/sec.
Al-Jumaily et al. (2013) → NER for Arabic text mining, no details on performance given	

Approach

arbitrary text

NE recognizer

as a combination of several
multi-layer finite state transducers
having different tolerance levels

voter

named entities
in text

connection to knowledge graph(s)
[instance labels, types, …]

access to language information
[word types, flections, …]

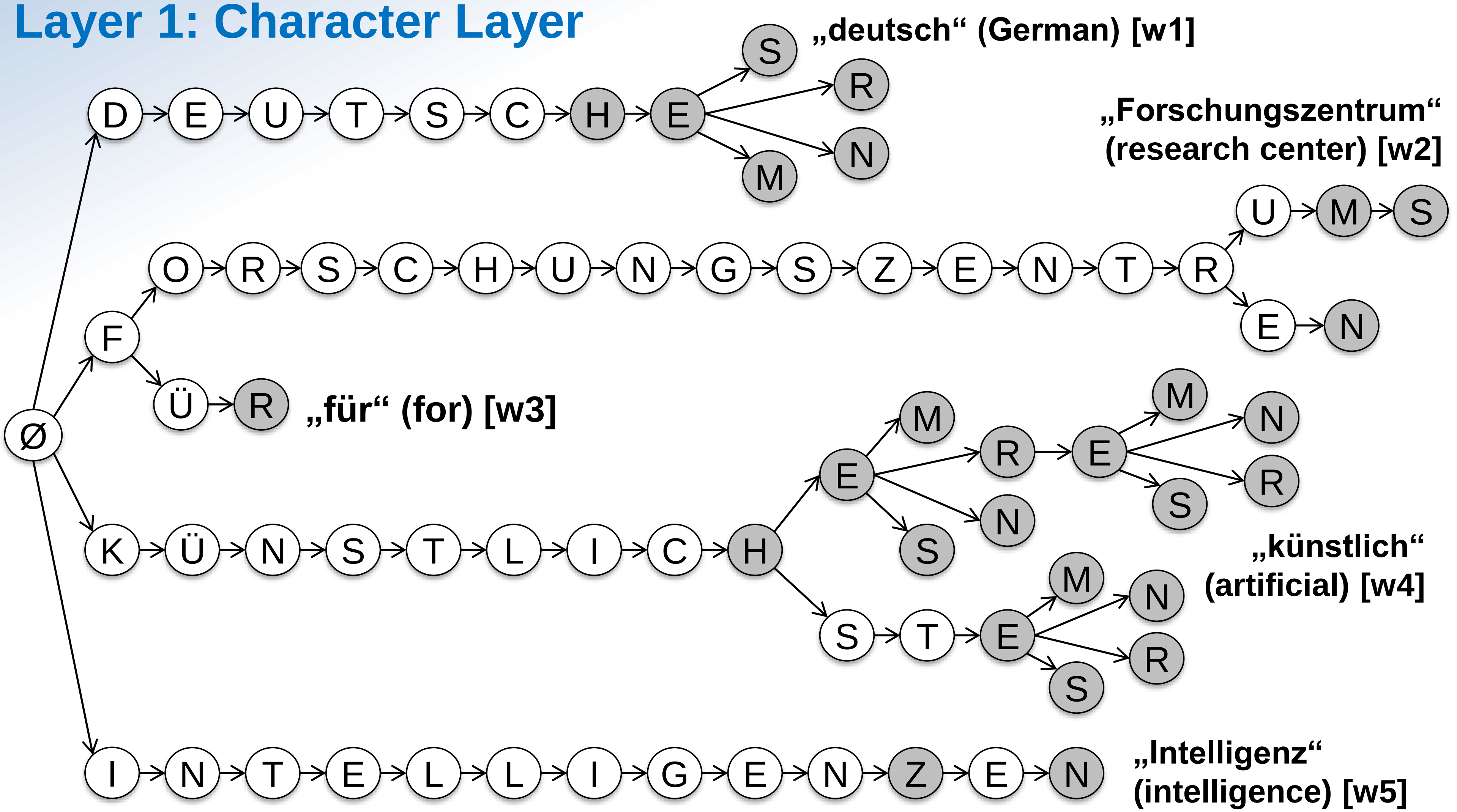
DBpedia

Wiktionary
[ˈvɪkʃəˌnɛʀi], *n*
Das freie Wörterbuch

LT Morphy

Multi-Layer FST with High Tolerance

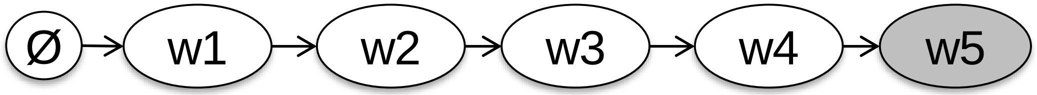
Layer 1: Character Layer



Input:

`Deutsches_ Forschungszentrum_ für_Künstliche_ Intelligenz`

Layer 2: Word Layer

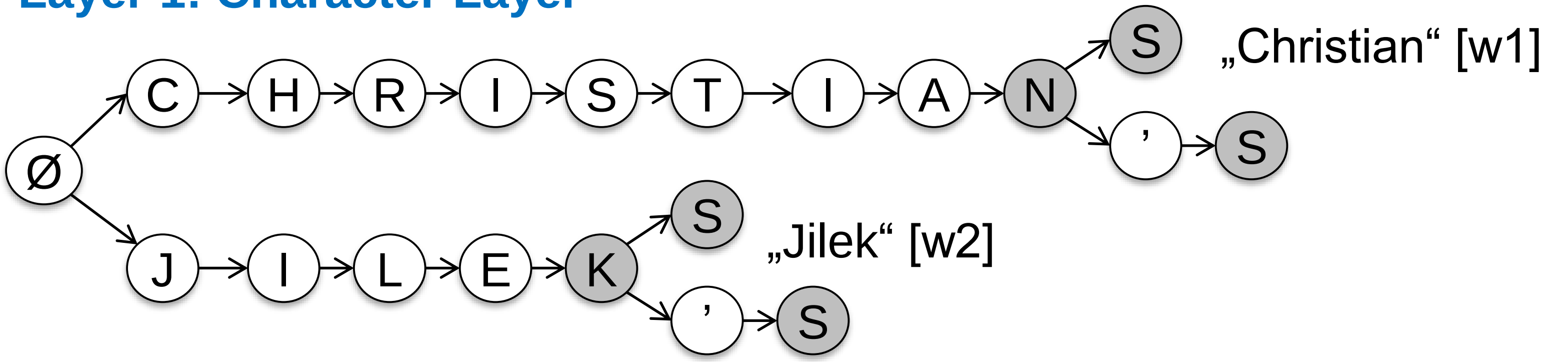


„Deutsches Forschungszentrum für Künstliche Intelligenz“ (German Research Center for Artificial Intelligence)

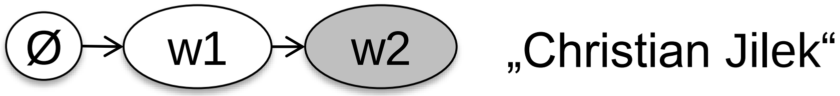
Multi-Layer FST with Low Tolerance

Input: `Christian_Jilek`

Layer 1: Character Layer



Layer 2: Word Layer



Evaluation Setting

- idea:
  - use the German Wikipedia as a large set of texts written by different people
  - use DBpedia types to decide whether to apply low or high inflection tolerance
  - use Wikipedia annotations as a „silver standard“
    - term used (often inflected form) manually annotated with its article name (often basic form)

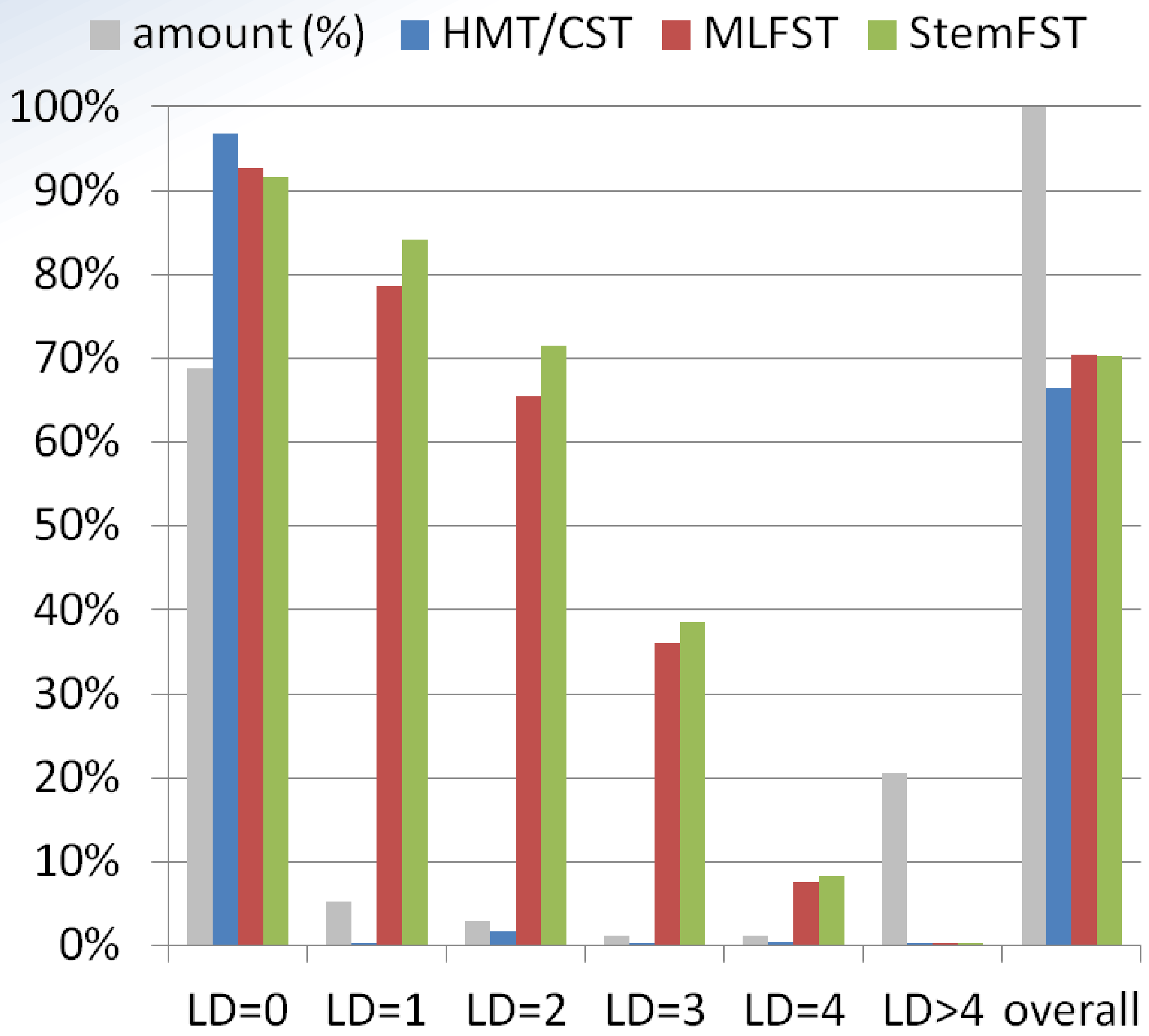
```
[[ Haus | Häuser ]]
[[ Junktor | Junktoren ]]
```

- problems:
  - independent term-links-combinations `[[ Eton | hometown ]]`
  - adjective-noun-combinations `[[ Entscheidbarkeit | entscheidbar ]]`

    → use Levenshtein distance (LD) to identify samples (typically LD<=4)

  - ambiguities (e.g. >1000 instances of „Jewish Cemetery“)
  - terms not annotated in „their own“ article (e.g. „Berlin“ in article about „Berlin“)
- benefit: 3.9M articles having 50.4M annotations

Results: Recall

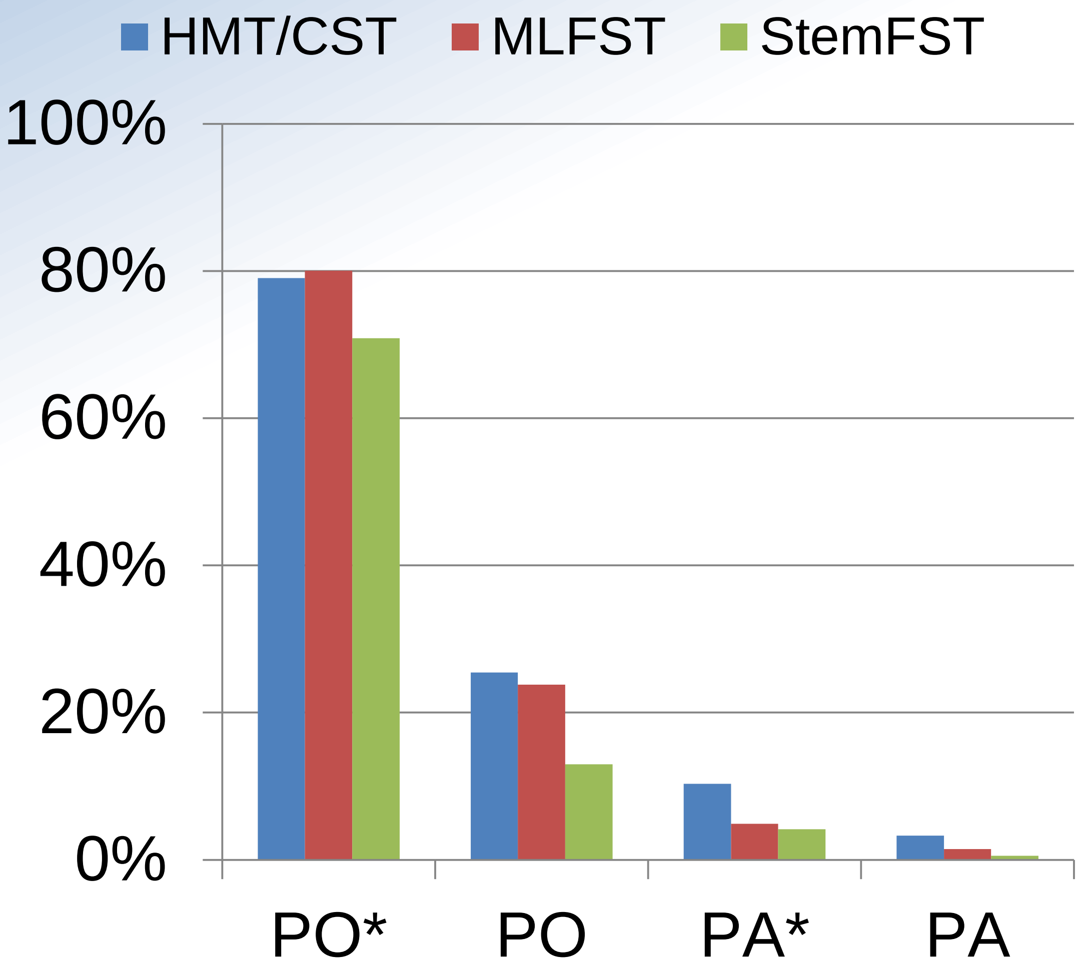


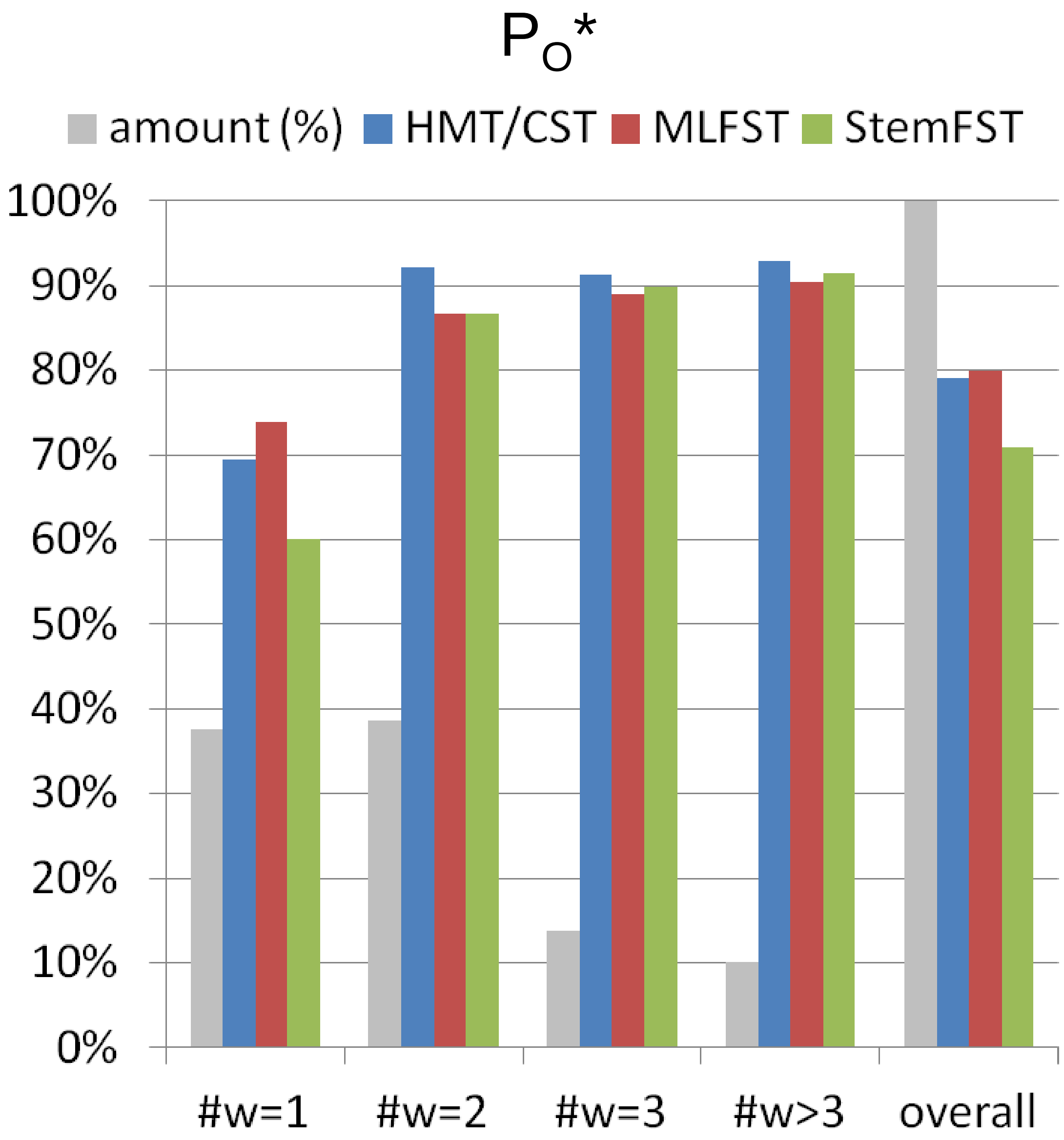
Results: Measuring Precision

„A commercial personal information management tool is used in the project.“

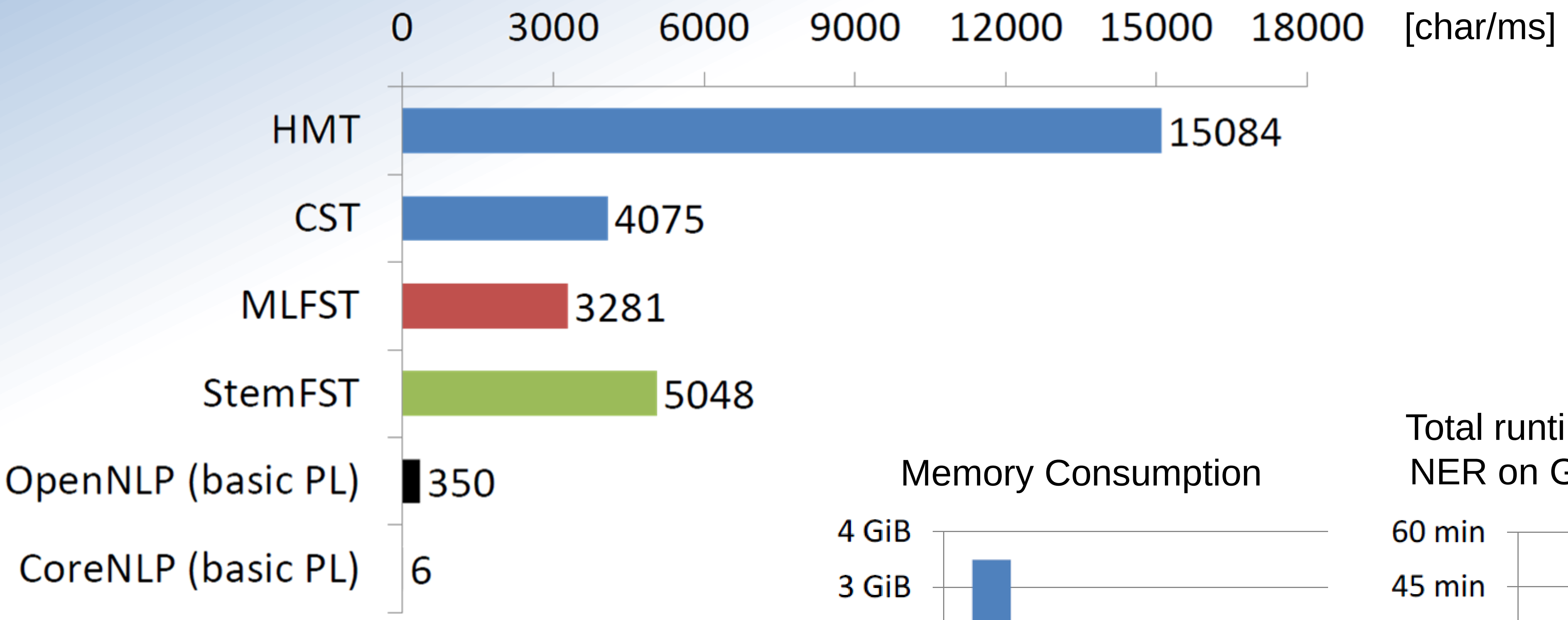
P_O^*	only overlapping terms as false positives, ambiguities disregarded
P_O	only overlapping terms including ambiguities as false positives
P_A^*	all other terms as false positives, ambiguities disregarded
P_A	all other terms including ambiguities as false positives

Results: Precision

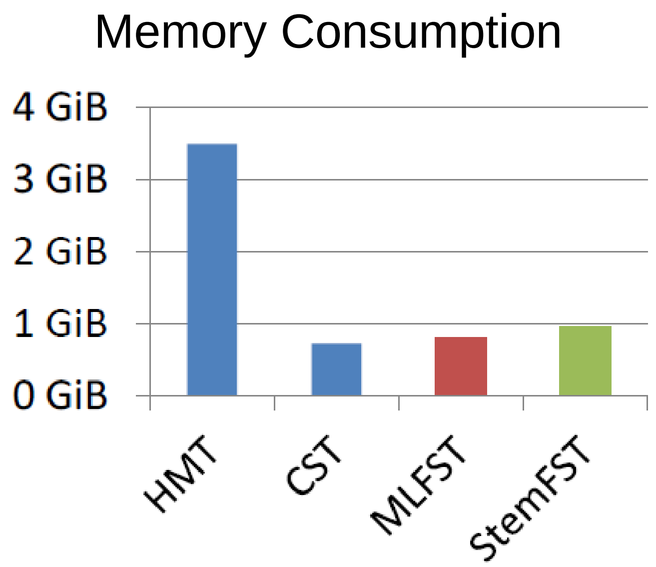


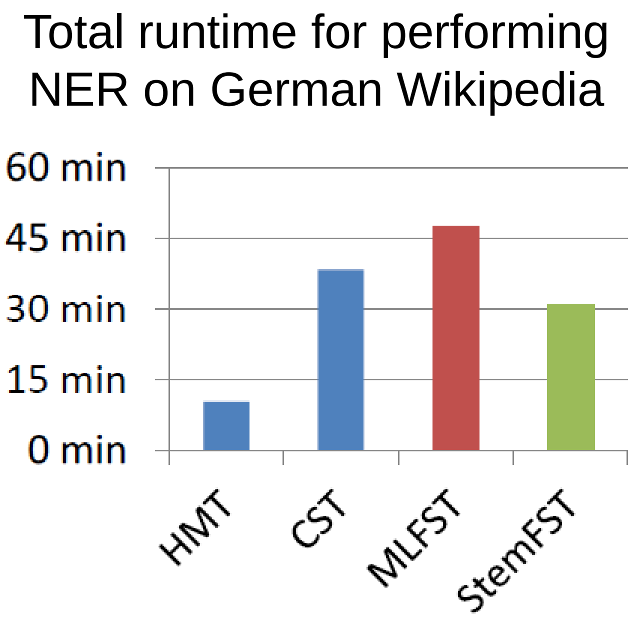


Results: Processing Speed & Memory Consumption



basic PL:
- tokenizer
- sentence splitter
- POS tagger





Conclusion

- presented inflection-tolerant and real-time capable OB NER approach based on
  - Trie-based string matching Aho & Corasick (1975)
  - finite state cascades Abney (1996)
  - exhaustive inflection listing Savary & Piskorski (2010)
  - exploiting ontological background information
- comparably fast as available high speed methods
- outperforming them in recognizing terms that lexically vary slightly (e.g. inflection)
- narrowing the gap to more sophisticated but slower NLP pipelines without losing too much runtime performance

Outlook

- incorporate disambiguation mechanisms (exploiting user context)
- add more layers to scan for patterns (ToDos, appointments, Hearst patterns, …)
- improve language capabilities (rules, heuristics, multi-language support, …)
- incorporate StemFST into MLFST for multi-word terms (slightly better precision)

Selected References

Abney, S.: **Partial parsing via nite-state cascades**. *Natural Language Engineering* 2(4), 337-344 (1996)

Aho, A.V., Corasick, M.J.: **Efficient string matching: an aid to bibliographic search**. *Communications of the ACM* 18(6), 333-340 (1975)

Al-Jumaily, H., Martnez, P., Martnez-Fernandez, J.L., Van der Goot, E.: **A real time named entity recognition system for arabic text mining**. *Language Resources and Evaluation* 46(4), 543-563 (2012)

Al-Rfou, R., Skiena, S.: **Speedread: A fast named entity recognition pipeline**. *Proc. 24th Int'l Conf. on Computational Linguistics (COLING 2012)* pp. 51-66 (2012)

Dey, A., Prukayastha, B.S.: **Named entity recognition using gazetteer method and n-gram technique for an inectional language: A hybrid approach**. *Int'l Journal of Computer Applications* 84(9) (2013)

Dlugolinsky, S., Nguyen, G., Laclavk, M., Seleng, M.: **Character gazetteer for named entity recognition with linear matching complexity**. In: *3rd World Cong. on Information and Communication Technologies (WICT)*. pp. 361-365. IEEE (2013)

Mendes, P.N., Jakob, M., Garca-Silva, A., Bizer, C.: **DBpedia spotlight: shedding light on the web of documents**. In: *Proc. of the 7th Int'l Conf. on Semantic Systems (I-Semantics)*. pp. 1-8. ACM (2011)

Nguyen, G., Dlugolinsky, S., Laclavik, M., Seleng, M., Tran, V.: **Next improvement towards linear named entity recognition using character gazetteers**. In: *Advanced Computational Methods for Knowledge Engineering*, pp. 255-265. Springer (2014)

Nielsen, J.: **Usability Engineering**. Morgan Kaufmann (1993)

Savary, A., Piskorski, J.: **Lexicons and grammars for named entity annotation in the national corpus of polish**. In: *18th Int'l Conf. Intelligent Information Systems*, pp. 141-154 (2010)

Thanks for your attention! ☺

Parts of this work were funded by the DFG in the SPP on Intentional Forgetting in Organizations.