UNIVERSITÄT DUISBURG ESSEN
Open-Minded

LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text

Tobias Horsmann, Torsten Zesch

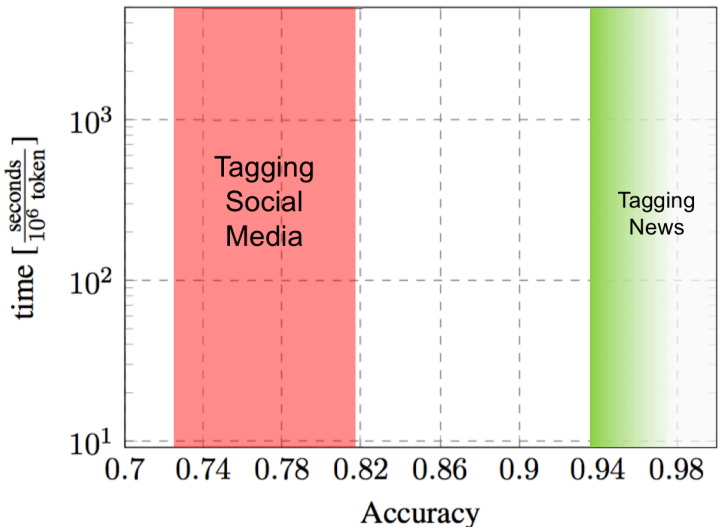
Language Technology Lab
University of Duisburg-Essen
Germany

LANGUAGE TECHNOLOGY LAB

Motivation

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Off-the-shelve PoS tagger perform poorly on social media text



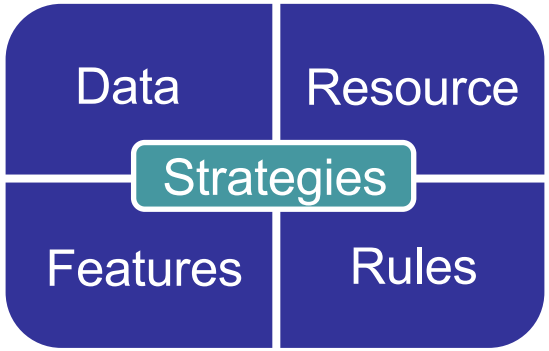
Problem:
How to improve tagging accuracy on (German) social media text

Empiri Shared Task

- Annotated data set (13k token) of two sub-domains
  - Computer Mediated Communication (CMC)
  - Web text (Web)
- Extended STTS tagset (54 + 18 new tags)

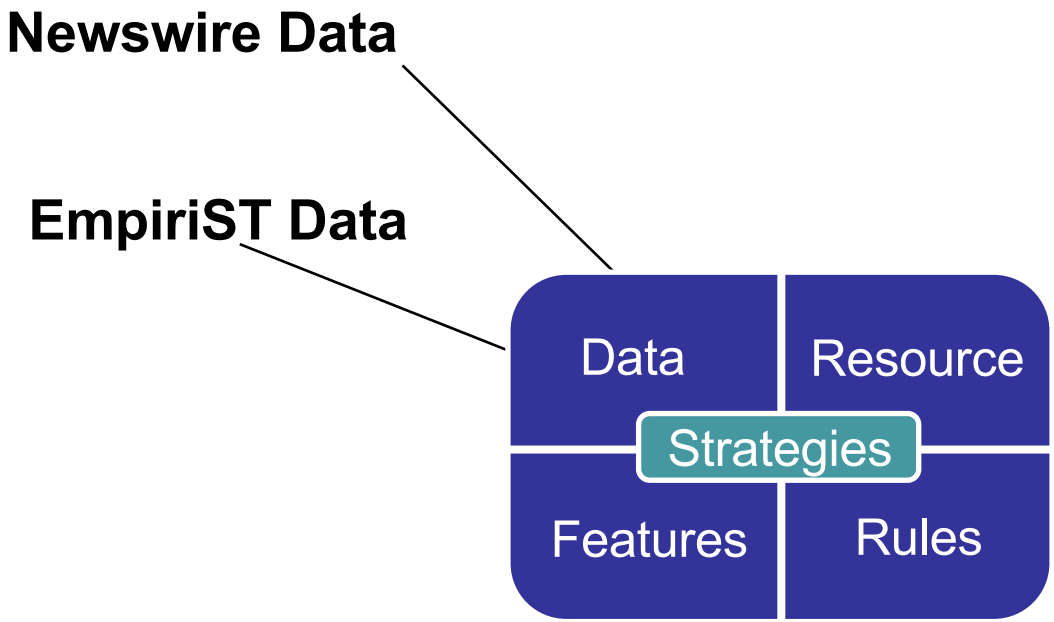
Adaptation Strategies

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

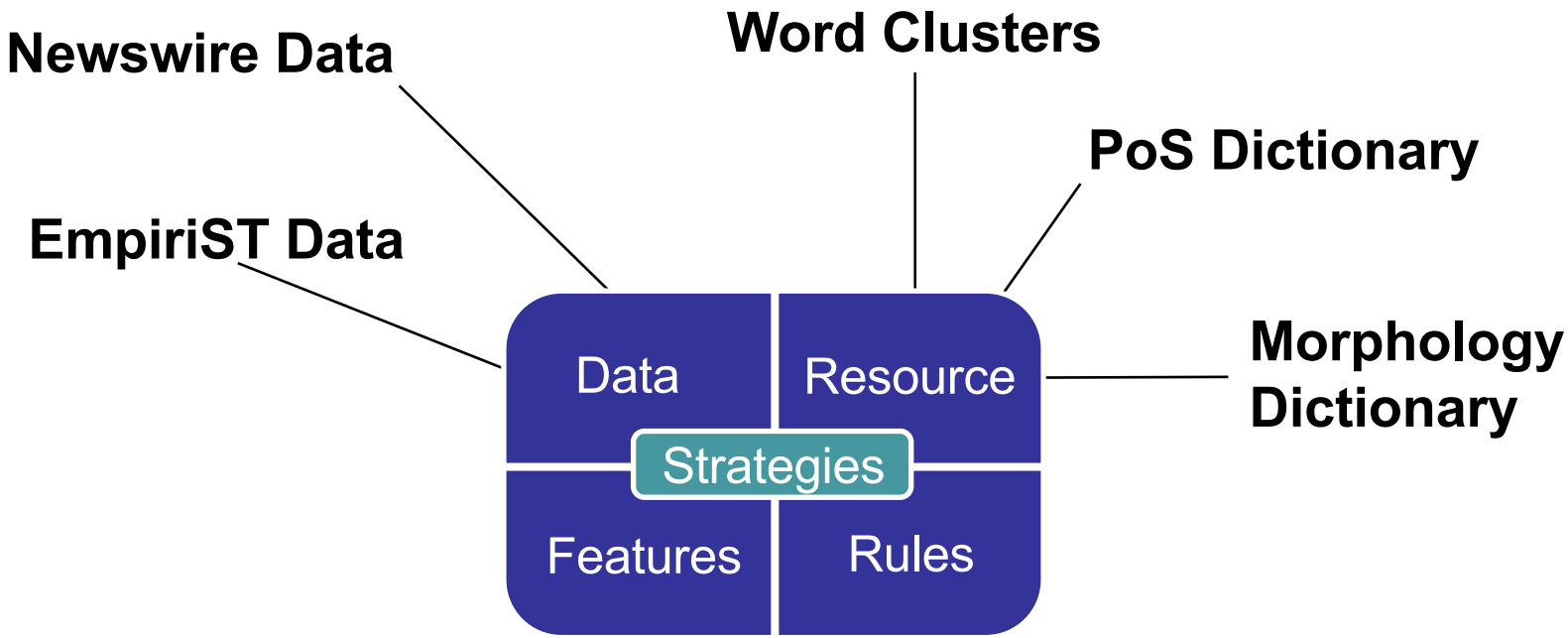


Adaptation Strategies

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

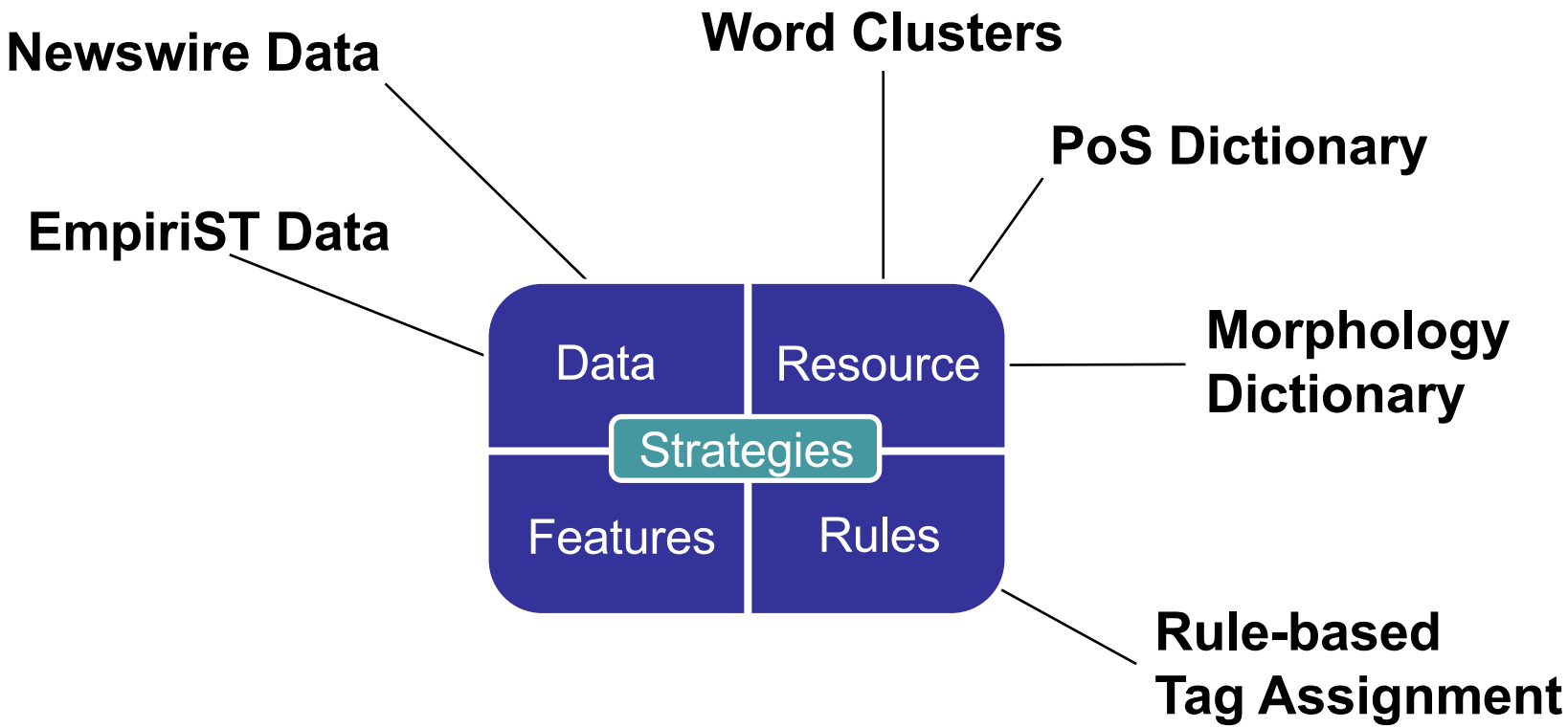


Adaptation Strategies

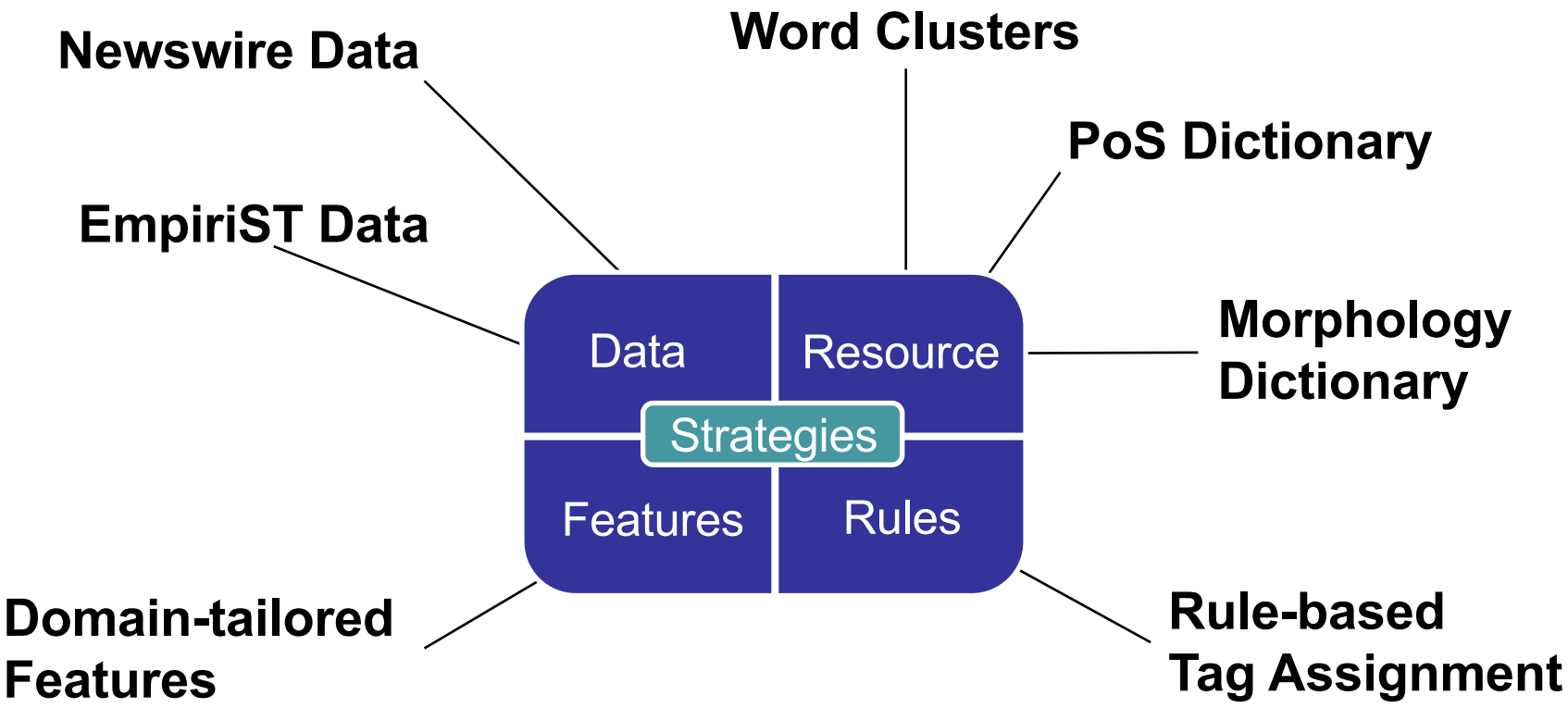


Adaptation Strategies

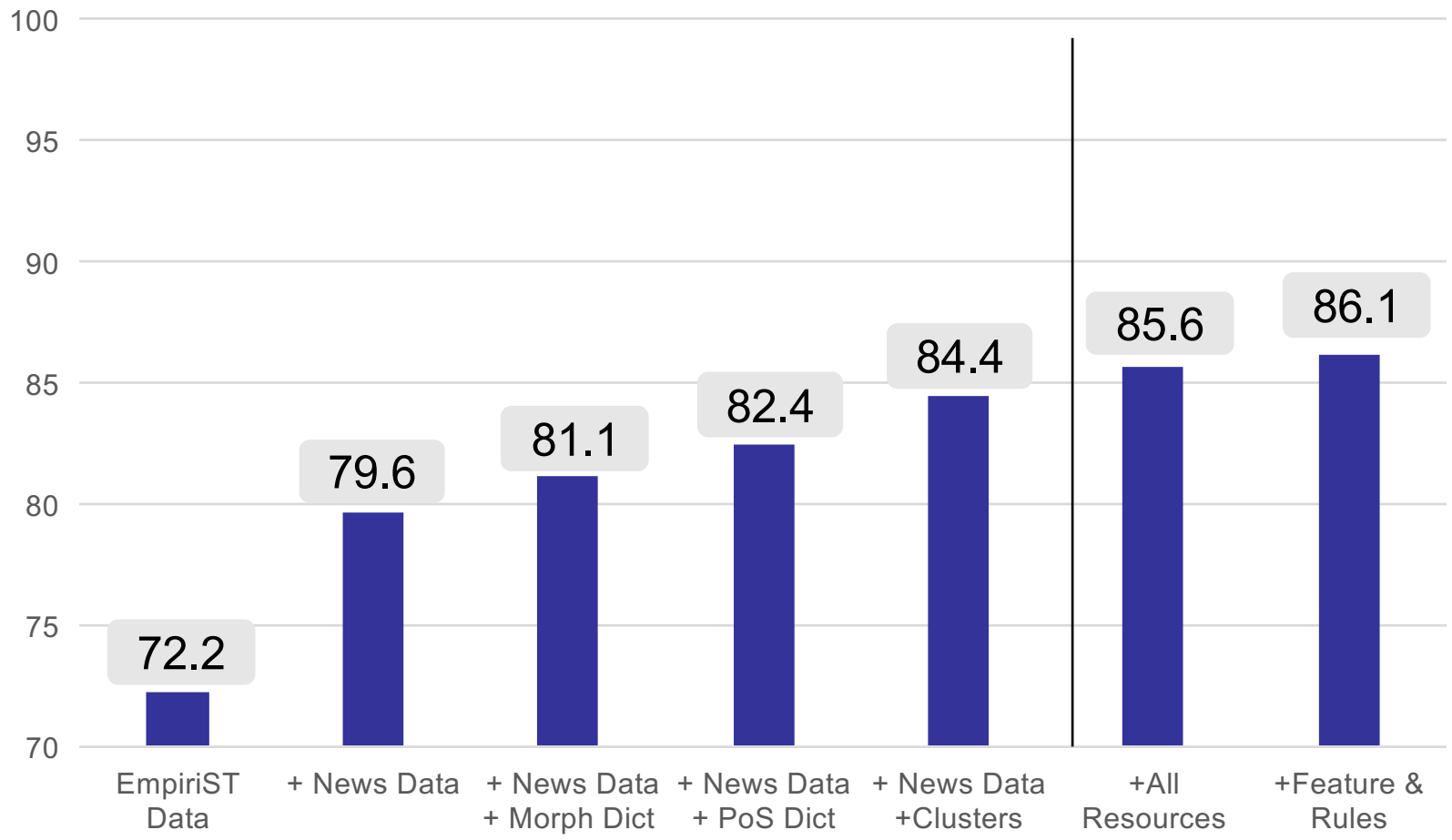
UNIVERSITÄT DUISBURG ESSEN
Open-Minded



Adaptation Strategies



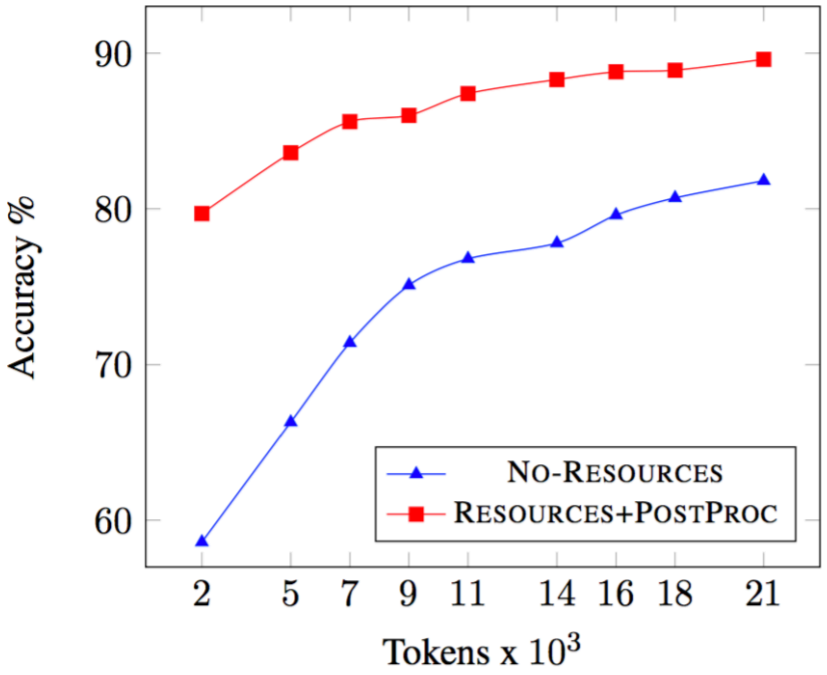
Results on EmpiriST Test Data



Do we (really) need more data?

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Learning Curve on EmpiriST Train+Test Data



Many low frequent (new) PoS tags

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

New Empiri PoS tags	Occurrence in Training Data	Accuracy (%) Test Data
EMOASC	115	97.2
PTKMA	103	20.0
PTKIFG	99	8.3
AKW	49	81.7
HST	46	97.6
ADR	35	54.2
PTKMWL	28	0.0
EMOIMG	22	90.5
URL	18	76.2
VVPPER	7	66.7
VAPPER	4	25.0
DM	3	0.0
VMPPER	1	0.0
ADVART	1	0.0
KOUSPPER	1	0.0
ONO	1	0.0
PPERPPER	1	0.0
EML	0	0.0

18 PoS tags have been newly introduced

Training Data

- 9 tags occur less than 10 times
- 1 tag does not even occur once

Test Data

- 8 tags have an accuracy of zero
- 11 tag have an accuracy below 50%

Conclusion

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Newswire tagging accuracy still out of reach

Resources (especially clusters) improve accuracy a lot

Low frequent tags hard to learn

→ devide into basic tagset / extended tagset

→ tackle in separate steps

→ select data for manual annotation to cover interesting parts