EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann
www.aiphes.tu-darmstadt.de

https://github.com/AIPHES

Introduction EmpiriST

Scenario:	Process **German web** data **and** computer mediated communication data (**social media**, CMC)
Training Data:	Web: 5,109 CMC: 6,034 Tokens + POS tags
Task:	1. Tokenization 2. POS tagging

Tokenization

1. Conservative splits at Unicode category change positions, e.g. 'poster A0' = 'Ll...ZsLuNd' => 'poster', ' ', 'A', '0'
2. Lookahead merge list:
   - Compiled from Wikipedia
   - Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
[1-2]
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a@b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\|\]\)\(DP+)\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

POS Parsing

Adapted *GermaNER* (Benikova et al.) Sequence Labelling Classificator based on CRFs and features:

1. **char n-grams**: n={1,2,3} from start and end of word
2. **word lists**, such as English Vocabulary, and manually compiled onomatopoeia (partially from internetslang.com)
3. **similar Words** from an automatically created distributional thesaurus (DT, 4 expansions per token)
4. **topic clusters**: based on LDA clustering on DT pseudo documents, i.e. one feature vector per token
5. simple **shallow features** such as word position and casing. Reuse regex rulesets from tokenization
6. **Unicode categories** of characters

Tiger was added for training, where simple STTS tag transformation rules were applied

Results

Tokenization:

Genre	*Rec*	*Prec*	F_1	*Rank*
CMC	99.30	98.62	98.96	2
WEB	99.63	99.89	99.76	2

POS Tagging:

Genre	*Acc*	*Rank*
CMC	84.22	5
CMC (STTS Map)	87.10	2
Web	93.27	2
Web (STTS Map)	94.30	2

Distribution of new labels in test data:

Tag	CMC	Web
ONO	2	
DM	6	
PTKIFG	72	61
PTKMA	74	11
PTKMWL	10	14
VVPPER	6	
VAPPER	4	
KOUSPPER	1	1
PPERPPER	1	1
ADVART	3	
EMOASC	71	
EMOIMG	63	
AKW	60	
HST	42	
ADR	48	
URL	16	
EML		1

Error Analysis

Common Tokenization Errors:

- Rules are underspecified

* Eingetr. Lebenspartnersch.	* Eingetr * .
* die * Feststellung der * 1. Teil meiner	* dieFeststellung , war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann

aiphes.tu-darmstadt.de

https://github.com/AIPHES

- Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
[1-2]
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a@b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\[\]\)\(DP•)\1{0,}-?'?[:;8B=]
# reject ]:
- \} ?:

...
```

Common Tokenization Errors:

- Rules are underspecified

* Eingetr. Lebenspartnersch.	* Eingetr * .
* die * Feststellung	* dieFeststellung ,
der * 1. Teil meiner	war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

EmpiriST: AIPHES Robust Tokenization for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT
AIPHES

...ith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann

...b.com/AIPHES

Introduction EmpiriST

Scenario: Process **German web** data **and** computer mediated communication data (**social media**, CMC)

Training Data: Web: 5,109
CMC: 6,034
Tokens + POS tags

Task:
1. Tokenization
2. POS tagging

Results

Tokenization:

Genre	*Rec*	*Prec*	F_1	*Rank*
CMC	99.30	98.62	98.96	2
WEB	99.63	99.89	99.76	2

POS Tagging:

Genre	*Acc*	*Rank*
CMC	84.22	5
CMC (STTS Map)	87.10	2
Web	93.27	2
Web (STTS Map)	94.30	2

Distribution of new labels in test data:

Tag	CMC	Web
ONO	2	
DM	6	
PTKIFG	72	61
PTKMA	74	11
PTKMWL	10	14
VVPPER	6	
VAPPER	4	
KOUSPPER	1	1
PPERPPER	1	1
ADVART	3	
EMOASC	71	
EMOMMG	63	
AKW	60	
HST	42	
ADR	48	
URL	16	
EML		1

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

Backlinks
:
* [[sec:verschl]]
Navigation
Passwort-
generator

...biguous phenomena (e.g. dates)

* 1.3.
Kekse

* 1.
* 3.

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

Tokenization

1. Conservative splits at Unicode category change positions, e.g. ‘poster A0’ = ‘Ll...ZsLuNd’
   => ‘poster’, ‘ ’, ‘A’, ‘0’
2. Lookahead merge list:
   - Compiled from Wikipedia
   - Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
Ü-Ei
altgriech.
24/7
2B~not2B
a-z
a.k.a.
>_<
:-}
X8-{}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a@b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\[\]\)\(DP*)\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

POS Parsing

Adapted *GermaNER* (Benikova et al.) Sequence Labelling Classificator based on CRFs and features:

1. **char n-grams**: n={1,2,3} from start and end of word
2. **word lists**, such as English Vocabulary, and manually compiled onomatopoeia (partially from internetslang.com)
3. **similar Words** from an automatically created distributional thesaurus (DT, 4 expansions per token)
4. **topic clusters**: based on LDA clustering on DT pseudo documents, i.e. one feature vector per token
5. simple **shallow features** such as word position and casing. Reuse regex rulesets from tokenization
6. **Unicode categories** of characters

Tiger was added for training, where simple STTS tag transformation rules were applied

- Freely available, Open So
- References:
  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

EmpiriST: AIPHES Robust Tokenizat… and POS-Tagging for Different G…

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judit…
www.aiphes.tu-darmstadt.de

https…

Introduction EmpiriST

Scenario:	Process **German web** data **and** computer mediated communication data (**social media**, CMC)
Training Data:	Web: 5,109 CMC: 6,034 Tokens + POS tags
Task:	1. Tokenization 2. POS tagging

Tokenization

1. Conservative splits at Unicode category change positions, e.g. 'poster A0' = 'Ll...ZsLuNd' => 'poster', ' ', 'A', '0'
2. Lookahead merge list:
   - Compiled from Wikipedia
   - Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
[1-9]
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a8b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\|\]\)\){DP+}\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

Common Tokenization…

- Rules are underspecif…

* Eingetr. Lebenspartnersch.	* Ein… * .
* die * Feststellung	* dieFeststel… ,
der * 1. Teil meiner	war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

Results

Tokenization:

Genre	*Rec*	*Prec*	F_1	*Rank*
CMC	99.30	98.62	98.96	2
WEB	99.63	99.89	99.76	2

POS Tagging:

Genre	*Acc*	*Rank*
CMC	84.22	5
CMC (STTS Map)	87.10	2
Web	93.27	2
Web (STTS Map)	94.30	2

Distribution of new labels in …ata:

Tag	CMC	Web
ONO	2	
DM	6	
PTKIFG	72	61
PTKMA	74	11
PTKMWL	10	14
VVPPER	6	

7. unknown…
8. unknown foreign…
9. error due to abbreviation
10. incorrect punctuation tag 20 12.3
11. other 26 16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann
www.aiphes.tu-darmstadt.de

https://github.com/AIPHES

Error Analysis

Common Tokenization Errors:

- Rules are underspecified

* Eingetr. Lebenspartnersch.	* Eingetr * .
* die * Feststellung	* dieFeststellung ,
der * 1. Teil meiner	war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Common Ta

Error class
1. missed extende
2. incorrectly assign
3. confusion of functi
4. mistagged *NN* due t
5. mistagged *NE* as *NN*
6. mistagged *NE* as other
7. unknown emoticon
8. unknown foreign langu
9. error due to abbreviati
10. incorrect punctuation
11. other

	$*	AD*	AKW*	AP*
$*	68	0	0	1
AD*	2	49	1	9
AKW*	0	1	0	0
AP*	2	0	0	4
AR*	0	1	0	0
EML*	0	0	0	0
EMO*	1	0	0	0
N*	1	25	2	
P*	0	3	0	
PPER*	0	0	0	
PT*	0	34		
V*	0			
other				

…German
…earch (DIPF)
…program.

EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann
www.aiphes.tu-darmstadt.de

https://github.com/AIPHES

Introduction EmpiriST

Scenario:	Process **German web** data **and** computer mediated communication data (**social media**, CMC)
Training Data:	Web: 5,109 CMC: 6,034 Tokens + POS tags
Task:	1. Tokenization 2. POS tagging

Tokenization

1. Conservative splits at Unico
   category change positions,
   'poster A0' = 'Ll...ZsLuNd'
   => 'poster', ' ', 'A', '0'
2. Lookahead merge list:
   - Compiled from Wikipe
   - Contains abbreviations

```
## lookahead-list.txt
C-Jugend
[-Ei
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expres
   - Positive Merge rules (
   - Negative Merge reject

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a@b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\|\]\)\(DP+)\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

POS Parsing

Adapted *GermaNER* (Benikova et al.)
Sequence Labelling Classificato
CRFs and features:

1. **char n-gram**
   and end
2. **wor**

Further

- Freely available, Open Source, Permissive Apache V2 Li
- References:
  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. automatic linguistic annotation of computer-mediated communication, social medi *10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German In *International Conference of the German Society for Computational Linguistics and Language Technolo* Germany.

rs:

ed

```
* Eingetr
* .
```

ch.

```
* dieFeststellung
  ,
  war
* der1
* .
  Teil
```

specified

```
Backlinks
:
* [[sec:verschl]]
Navigation
Passwort-
generator
```

(e.g. dates)

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

		$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	
PT*	0	34	0	5	0	0	0	1	7	0	16		
V*	0	16	3	4	1	0	0	30	1	0	0		
other	11	10	5	9	1	0	1	28	9	1			

EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann
www.aiphes.tu-darmstadt.de

https://github.com/AIPHES

Introduction EmpiriST

Scenario:	Process **German web** data **and** computer mediated communication data (**social media**, CMC)
Training Data:	Web: 5,109 CMC: 6,034 Tokens + POS tags
Task:	1. Tokenization 2. POS tagging

Tokenization

1. Conservative splits at Unicode category change positions, e.g. 'poster A0' = 'Ll...ZsLuNd' => 'poster', ' ', 'A', '0'
2. Lookahead merge list:
   - Compiled from Wikipedia
   - Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
[1-2]
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a@b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\|\]\)\(DP+)\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

POS Parsing

Adapted *GermaNER* (Benikova et al.) Sequence Labelling Classificator based on CRFs and features:

1. **char n-grams**: n={1,2,3} from start and end of word
2. **word lists**, such as English Vocabulary, and manually compiled onomatopoeia (partially from internetslang.com)
3. **similar Words** from an automatically created distributional thesaurus (DT, 4 expansions per token)
4. **topic clusters**: based on LDA clustering on DT pseudo documents, i.e. one feature vector per token
5. simple **shallow features** such as word position and casing. Reuse regex rulesets from tokenization
6. **Unicode categories** of characters

Tiger was added for training, where simple STTS tag transformation rules were applied

Results

Tokenization:

Genre	*Rec*	*Prec*	F_1	*Rank*
CMC	99.30	98.62	98.96	2
WEB	99.63	99.89	99.76	2

POS Tagging:

Genre	*Acc*	*Rank*
CMC	84.22	5
CMC (STTS Map)	87.10	2
Web	93.27	2
Web (STTS Map)	94.30	2

Distribution of new labels in test data:

Tag	CMC	Web
ONO	2	
DM	6	
PTKIFG	72	61
PTKMA	74	11
PTKMWL	10	14
VVPPER	6	
VAPPER	4	
KOUSPPER	1	1
PPERPPER	1	1
ADVART	3	
EMOASC	71	
EMOIMG	63	
AKW	60	
HST	42	
ADR	48	
URL	16	
EML		1

Error Analysis

Common Tokenization Errors:

- Rules are underspecified

* Eingetr. Lebenspartnersch.	* Eingetr * .
* die * Feststellung der * 1. Teil meiner	* dieFeststellung , war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.

EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

https://github.com/AIPHES

Training Data:	Web: 5,109 CMC: 6,034 Tokens + POS tags
Task:	1. Tokenization 2. POS tagging

Tokenization

1. Conservative splits at Unicode category change positions, e.g. 'poster A0' = 'Ll...ZsLuNd' => 'poster', ' ', 'A', '0'
2. Lookahead merge list:
   - Compiled from Wikipedia
   - Contains abbreviations, emoticons, etc.

```
## lookahead-list.txt
C-Jugend
[1-9]
altgriech.
24/7
2B-not2B
a-z
a.k.a.
>_<
:-)
X8-(}
...
```

3. Lookahead merge rules:
   - Contains regular expressions
   - Positive Merge rules (+)
   - Negative Merge reject rules (-)

```
## lookahead-rules.txt
# reject ) followed by ; globally
- \);

# email a8b.com
+ [\.+\w\-]+@([\w\-]+\.)+[\w]{2,6}

# reverse emoticons (-:
+ (\|\]\)\(DP+)\1{0,}-?'?[:;8B=]
# reject ):
- \) ?:

...
```

onomatopoeia (partially from internetslang.com)

3. **similar Words** from an automatically created distributional thesaurus (DT, 4 expansions per token)
4. **topic clusters**: based on LDA clustering on DT pseudo documents, i.e. one feature vector per token
5. simple **shallow features** such as word position and casing. Reuse regex rulesets from tokenization
6. **Unicode categories** of characters

Tiger was added for training, where simple STTS tag transformation rules were applied

	Acc.	Rank
CMC	84.22	5
CMC (STTS Map)	87.10	2
Web	93.27	2
Web (STTS Map)	94.30	2

Distribution of new labels in test data:

Tag	CMC	Web
ONO	2	
DM	6	
PTKIFG	72	61
PTKMA	74	11
PTKMWL	10	14
VVPPER	6	
VAPPER	4	
KOUSPPER	1	1
PPERPPER	1	1
ADVART	3	
EMOASC	71	
EMOIMG	63	
AKW	60	
HST	42	
ADR	48	
URL	16	
EML		1

Error Analysis

Common Tokenization Errors:

- Rules are underspecified

* Eingetr. Lebenspartnersch. * die * Feststellung der * 1. Teil meiner	* Eingetr * . * dieFeststellung , war * der1 * . Teil

- Rules are overspecified

Backlinks : * [[* sec : verschl	Backlinks : * [[sec:verschl]] Navigation Passwort- generator

- Ambiguous phenomena (e.g. dates)

* 1.3. Kekse	* 1. * 3.

Common Tagging Errors:

Error class	count	(%)
1. missed extended tagset	28	17.6
2. incorrectly assigned new tag	4	2.5
3. confusion of function word tags	22	13.6
4. mistagged *NN* due to lower case	23	14.4
5. mistagged *NE* as *NN* & vice-versa	20	12.5
6. mistagged *NE* as other	12	10.0
7. unknown emoticon	1	0.6
8. unknown foreign language word	2	1.3
9. error due to abbreviation	2	1.3
10. incorrect punctuation tag	20	12.5
11. other	26	16.3

	$*	AD*	AKW*	AP*	AR*	EML*	EMO*	N*	P*	PPER*	PT*	V*
$*	68	0	0	1	1	0	0	1	0	1	0	1
AD*	2	49	1	9	0	1	0	81	21	3	159	21	34
AKW*	0	1	0	0	0	0	0	1	0	0	0	0	3
AP*	2	0	0	4	0	0	0	0	1	0	6	2	10
AR*	0	1	0	0	0	0	0	0	22	0	2	0	0
EML*	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO*	1	0	0	0	0	0	0	0	0	0	0	0	0
N*	1	25	2	3	3	0	2	163	3	2	6	16	43
P*	0	3	0	1	10	0	0	4	37	4	1	1	10
PPER*	0	0	0	0	0	0	0	0	7	0	0	0	0
PT*	0	34	0	5	0	0	0	1	7	0	16	0	0
V*	0	16	3	4	1	0	0	30	1	0	0	137	7
other	11	10	5	9	1	0	1	28	9	1	5	1	25

Further Information

- Freely available, Open Source, Permissive Apache V2 License
- References:

  - Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication, social media and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, Berlin, Germany.
  - Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free open German named entity recognition tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL- 2015)*, Essen, Germany.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Infor- mation from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1 and by the German Institute for Educational Research (DIPF) under the KDSL program.