EmpiriST 2015:

A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora

Michael Beißwenger
(U Duisburg-Essen)

Sabine Bartsch
(TU Darmstadt)

Stefan Evert
(FAU Erlangen-Nürnberg)

Kay-Michael Würzner
(BBAW, Berlin)

10th Web as Corpus Workshop (WAC-X)

12 August 2016

EmpiriST 2015: Background and motivation

Project defined in the context of the DFG scientific network **Empirikom** (2010-2014, http://www.empirikom.net/):

"**Empirikom** **S**hared **T**ask on Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora"

Computer-mediated communication (CMC): written discourse in genres such as chats, blogs, forums, tweets, newsgroups, SMS, whatsapp, on „social network" sites and on wiki talk pages

- CMC discourse typically includes a range of deviations from the syntactic and orthographic norms of the written standard (often referred to as 'non-canonical' language or 'non-standard' writing)
- How can these phenomena be handled with NLP approaches and tools?

Goals of the task

- In the past years: increasing number of approaches for adapting NLP tools to CMC:
  - Ritter et al. 2011; Gimpel et al. 2011; Owoputi et al. 2015; Avontuur et al. 2012; Bartz et al. 2013; Neunerdt et al. 2013; Rehbein 2013; Rehbein et al. 2013; Horbach et al. 2014, 2015; Horsmann/Zesch 2015; Ljubesic et al., 2015; …
- **Goal of the task:** foster the further development and comparison of approaches in a competitive setting on basis of the same data sets (= community shared task)

Perspectives, focus and schedule

Two perspectives:

- **Linguistics** (corpus-based analysis of computer-mediated communication and linguistic variation)
- **Computational Linguistics** (building and analyzing large web corpora)

Focus of the task (= two substasks):

1. Tokenization
2. Part-of-speech tagging

Schedule:

- Conception and preparation of the task: 2014-15
- Trial data and annotation guidelines: October 2015
- Release of full task: February 2016
- System submission: May 2016

Research questions

Research questions:

1. To what extent can the performance of automatic tools for tokenization and PoS tagging of German CMC discourse be improved, using our gold standard for training or domain adaptation?

2. Can both genuine CMC corpora andWeb corpora (where CMC phenomena typically occur much less frequently) be processed by the same approaches and models, or do we need different tools for the two types of corpora?

The gold standard

> 20,000 tokens of manually annotated data (tokenized and tagged independently by 2-5 student annotators):

1. A genuine ***CMC data set*** with samples from several CMC genres (chats, tweets, blog comments, Wikipedia talk pages, whatsapp interactions)
   - 5,109 tokens of training data
   - 5,234 tokens of test data

2. A ***Web corpora data set*** of CC-licensed Web pages which represents the type of data found in large corpora crawled from the Web
   - 4,944 tokens of training data
   - 7,568 tokens of test data

   (Participants were allowed to use additional data sets for training)

Annotation guidelines

Tokenization:

- Guideline with detailed rules for handling CMC-specific tokenization issues (e.g., handling of speedwriting phenomena, irregular and creative use of whitespace and punctuation, emoticons and addressing terms)

PoS tagging:

- **STTS_IBK tagset:** based on the Stuttgart-Tübingen Tagset (STTS, Schiller et al. 1999) with additional tags for elements specific to CMC
- Detailed guideline (German & English)

Before the annotation of the task data sets, draft versions of both guidelines have been tested with student annotators.

New tags in the tagset *STTS_IBK*

PoS tag	Category	Examples
I. Tags for phenomena specific for CMC / social media discourse:		
EMOASC	ASCII emoticon	:-) :-(^^ O.O
EMOIMG	Graphic emoticon (emoji)	☺ ☹ ☻
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #Urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://tu-dortmund.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena typical for spontaneous spoken ('conceptually oral') language in colloquial registers:		
VVPPER	Tags for different types of colloquial contractions thate are frequent in CMC (APPRART already exists in STTS 1.0)	schreibste, machste
APPRART		vorm, überm, fürn
VMPPER		willste, darfste, musste
VAPPER		haste, biste, isses
KOUSPPER		wenns, weils, obse
PPERPPER		ichs, dus, ers
ADVART		son, sone
PTKIFG	Intensifier, focus and gradation particles	sehr schön, höchst eigenartig, nur sie, voll geil
PTKMA	Modal particles and downtoners	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTKMWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ...*with V2 clauses*
ONO	Onomatopoeia	boing, miau, zisch

Not repre-sented in STTS 1999 (except APPRART)

STTS 1999: included in ADV (‚adverb')

Table 2: Tagset extensions for CMC phenomena in STTS_IBK. More examples with context can be found in the detailed annotation guidelines on the EmpiriST Web site (available in German and English).

Availability

The gold standard and annotation guidelines are (and will remain) available on the shared task website:

EmpiriST 2015

Diese Site durchsuchen

Navigation
- Instructions & schedule
- Data sets
- **Gold standard**
- Annotation Guidelines
- Results
- Task Force / Contact
- Sitemap

GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / Social Media >

Gold standard

The EmpiriST 2015 gold standard comprises more than 20,000 tokens of CMC and Web corpora data that have been manually tokenized and annotated with PoS tags. It is divided into training and tests sets, which correspond exactly to the data used in the shared task. The gold standard is released under a Creative Commons CC BY-SA 3.0 licence.

- Download the gold standard: empirist_gold_standard.zip

If you use these data, please cite our task description paper:

Michael Beißwenger, Sabine Bartsch, Stefan Evert and Kay-Michael Würzner (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-med
EmpiriST Shared Task, pages

Data sizes

Training data:

- CMC subset: 5,109 toker
- Web corpora subset: 4,9

Test data:

- CMC subset: 5,234 toker
- Web corpora subset: 7,5

File formats

- Raw texts are provided a structure (postings or pa
- Manually tokenized texts Unix UTF-8 encoding. N whitespace and inserting
- Manually tagged texts a

EmpiriST 2015

Diese Site durchs

Navigation
- Instructions & schedule
- Data sets
- Gold standard
- **Annotation Guidelines**
- Results
- Task Force / Contact
- Sitemap

GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / Social Media >

Annotation Guidelines

The training data that will be provided as a gold standard have been manually tokenized and tagged according to the following guidelines:

- Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): **Tagset und Richtlinie für das Par Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation.** Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (*EmpiriST 2015*). (21 pa
  PDF (German): EmpiriST_Guideline-PoS.pdf
  PDF (English): EmpiriST_guideline-PoS(english).pdf (translated by Sabine Bartsch)
- Beißwenger, Michael; Bartsch, Sabine; Evert, Stefan; Würzner, Kay-Michael (2015): **Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation.** Guideline document from the Empi shared task on automatic linguistic annotation of internet-based communication (*EmpiriST 2015*). (29 pages).
  PDF (German): EmpiriST_Guideline-Tokenisierung.pdf
- **Ergänzungsdokument zu den Annotationsrichtlinien**: Additional instructions and examples for selected PoS catego

Credits (for support during the preparation stage)

Tools:

- Web-based corpus annotation platform CorA
  (Thanks to Stefanie Dipper & Marcel Bollmann, U Bochum)
- Web-based annotation tool MiniMarker (FAU Erlangen-Nürnberg)

People:

- Student annotators in Berlin, Darmstadt, Dortmund & Erlangen

Input on the conception of the annotation guidelines:

- Colleagues from the Empirikom network
- Colleagues from three workshops (2012–13) on adaptation of STTS to new genres (initiated by U. Heid, H. Zinsmeister, K. Beck)
- Colleagues from the CLARIN-D curation project ChatCorpus2CLARIN and at the FOLK corpus (spoken language) at IDS Mannheim

Financial support:

GSCL

- German Society for Computational Linguistics and Language Technology (GSCL)

Phase 1: Tokenization

15.02.–19.02.2016

Input:

Gehe
heute ins
Kino... :-)
freu

Output:

Gehe
heute
ins
Kino
...
:-
)
freu

Gold:

Gehe
heute
ins
Kino
...
:-)
*
freu
*

+ filler data

Phase 1: Tokenization

15.02.–19.02.2016

Sys: G e h e | h e u t e | i n s | K i n o | . . . | : - |) | * f r e u * |

Gold: G e h e | h e u t e | i n s | K i n o | . . . | : -) | * | f r e u | * |

Precision: $P = 7 / 8 = 87.5\%$

Recall: $R = 7 / 9 = 77.7\%$

F-Score: $F_1 = 2 \times P \times R / (P + R) =$ **82.4%**

(Jurish & Würzner 2013)

Phase 2: POS-Tagging

22.02.–26.02.2016

Input:

Gehe
heute
ins
Kino
...
:-)
*
freu
*

Output:

Gehe	VVFIN
heute	ADV
ins	APPRART
Kino	NN
...	$.
:-)	~~$(~~
*	~~XY~~
freu	~~XY~~
*	~~XY~~

Gold:

Gehe	VVFIN
heute	ADV
ins	APPRART
Kino	NN
...	$.
:-)	EMOASC
*	$(
freu	AKW
*	$(

Accuracy: **acc** = 5 / 9 = **55.6%** [best of up to 3 runs]

(Giesbrecht & Evert 2009)

Annotator agreement

tokenization
Web corpora

	AM	AS	DP	JM	LS
gold	99.56	99.74	99.70	99.78	99.93
AM		99.75	99.67	99.66	99.62
AS			99.88	99.89	99.80
DP				99.87	99.71
JM					99.73

tagging
Web corpora

	AM	AS	JM	LS
gold	92.64	96.15	95.49	91.77
AM		91.54	90.80	88.42
AS			93.04	89.51
JM				90.27

5th annotator:
95.9% vs. gold

Similar agreement on **CMC subset** (see paper)

Task participants

team ID	token	POS	remarks
AIPHES	✅	✅	CRF, rich feature set
bot.zen		✅	late submission, LSTM
COW	✅	✅	existing COW pipeline adapted
LTL-UDE	✅	✅	CRF, rich feature set
SoMaJo	✅		rule-based tokenizer
$WAGMOB	✅	✅	non-competitive
UdS		✅	HMM, additional training data
WASTE	✅		baseline (trained on EmpiriST)
Stanford Tagger	✅	✅	baseline (off the shelf)
TreeTagger	✅	✅	baseline (off the shelf)

Relaxed POS evaluation (STTS 1.0)

- Goal: comparison with existing baseline systems
- Define STTS 1.0 equivalents for new tags in gold standard
  - URL → XY
  - ADR → XY, NE
  - ONO → ITJ, VVFIN, VVIMP, VVINF
- But don't punish STTS_IBK systems
  - PTKIFG → ADV, ADJD, PTKMA, PTKMWL
  - PTKMA → ADV, ADJD, PTKIFG, PTKMWL
  - PTKMWL → ADV, ADJD, PTKIFG, PTKMA
- Extra rule for the Stanford Tagger
  - $(→ TRUNC

gold tag	these tags are also accepted
EMOIMG	XY ITJ EMOASC
AKW	VVFIN VVIMP VVINF VVIZU VAFIN VAIMP VAINF VMFIN VMINF
HST	XY
ADR	XY NE
URL	XY
EML	XY
VVPPER	VVFIN VVIMP VVINF
VMPPER	VMFIN VMINF
VAPPER	VAFIN VAIMP VAINF
KOUSPPER	KOUS
PPERPPER	PPER
ADVART	ART
PTKIFG	ADV ADJD PTKMA PTKMWL
PTKMA	ADV ADJD PTKIFG PTKMWL
PTKMWL	ADV ADJD PTKIFG PTKMA
DM	KOUS ADV
ONO	ITJ VVFIN VVIMP VVINF
ADV	PTKIFG PTKMA PTKMWL DM
KOUS	DM
PIDAT	PIAT

Results: Tokenization

team ID	CMC		Web		overall		
	F_1	rank	F_1	rank	F_1	rank	pdm
SoMaJo	99.54	1	99.60	3	99.57	1	**1**
AIPHES	98.96	2	99.76	2	99.36	2	**2**
COW	98.18	5	99.77	1	98.98	3	**3**
WASTE	98.47	4	99.42	4	98.95	4	**–**
LTL-UDE	98.58	3	99.22	8	98.90	5	**4**
$WAGMOB	97.83	6	99.39	5	98.61	6	**–**
Stanford Tagger	97.41	7	99.34	7	98.38	7	**–**
TreeTagger	94.96	8	99.36	6	97.16	8	**–**

Results: POS tagging

team ID	CMC		Web		overall		
	acc	rank	acc	rank	acc	rank	pdm
UdS-distributional$_2$	87.33	1	93.55	1	90.44	1	**1**
UdS-retrain$_2$	86.40	3	92.79	3	89.60	2	–
UdS-surface$_2$	86.45	2	92.43	4	89.44	3	–
LTL-UDE$_2$	86.07	4	92.10	5	89.09	4	**2**
AIPHES	84.22	7	93.27	2	88.75	5	**3**
bot.zen$_3$	85.42	5	90.63	8	88.03	6	–
COW	77.89	8	91.82	6	84.86	7	–
$WAGMOB	84.77	6	84.51	10	84.64	8	–
TreeTagger	73.21	9	91.75	7	82.48	9	–
Stanford Tagger	70.60	10	89.42	9	80.01	10	–

Results: POS tagging (STTS 1.0)

team ID	CMC		Web		overall		
	acc	rank	acc	rank	acc	rank	pdm
UdS-distributional$_2$	90.28	1	94.62	1	92.45	1	**1**
UdS-retrain$_2$	89.07	3	93.86	3	91.47	2	–
UdS-surface$_2$	89.28	2	93.50	4	91.39	3	–
LTL-UDE$_2$	88.84	4	93.12	5	90.98	4	**2**
AIPHES	87.10	6	94.30	2	90.70	5	**3**
bot.zen$_3$	87.47	5	91.74	9	89.61	6	–
COW	81.51	8	92.96	6	87.24	7	–
$WAGMOB	87.03	7	85.57	10	86.30	8	–
TreeTagger	76.81	9	92.89	7	84.85	9	–
Stanford Tagger	75.83	10	92.52	8	84.18	10	–

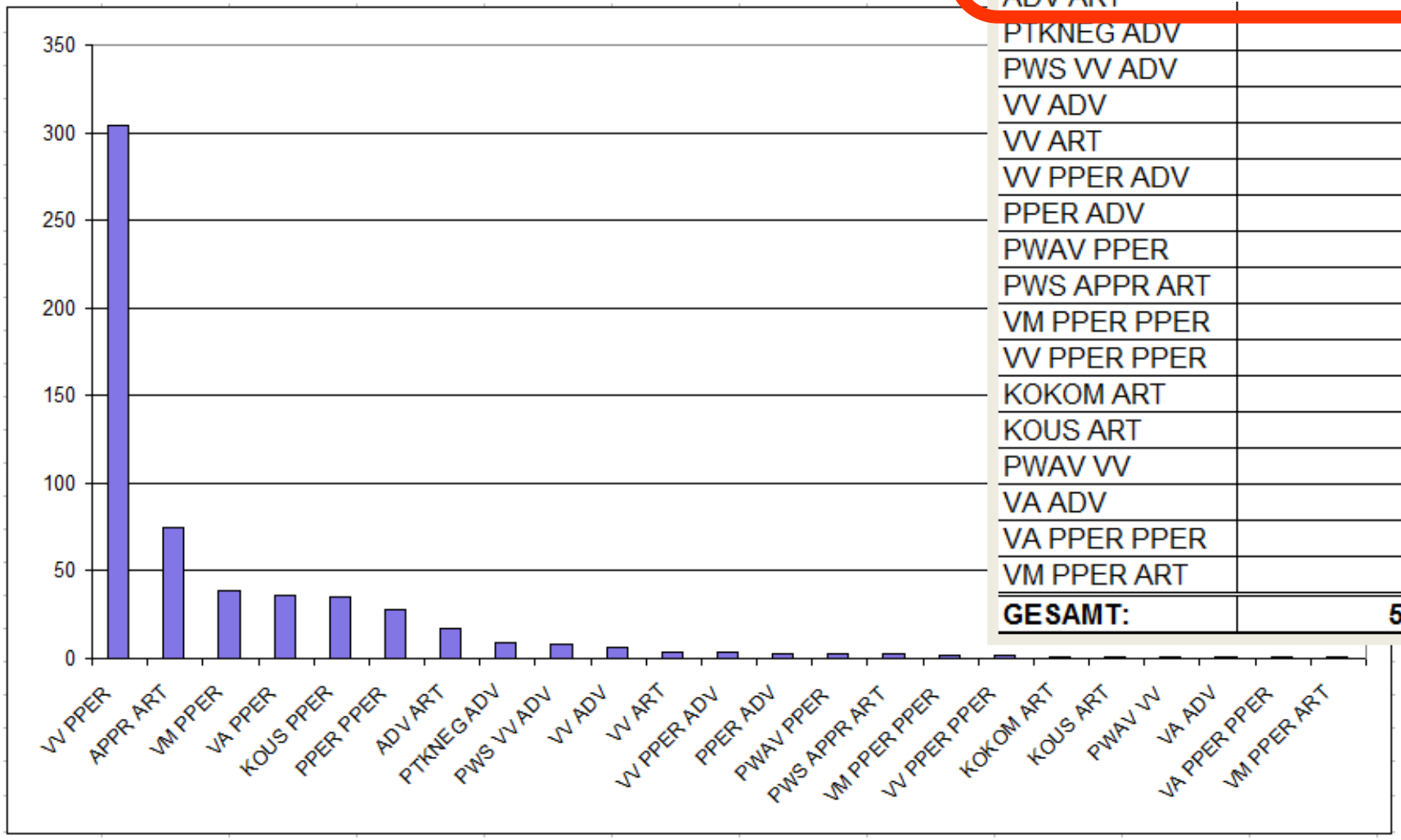
Summary & outlook

- Significant improvements over off-the-shelf baselines both for tokenization and POS tagging
  - not just because of tokenization guidelines and extended tag set
- Much larger improvement for CMC than for Web data
- Tokenization on par with human annotators → solved
- Tagging still has much room for improvement
  - CMC: 87.33% vs. 81.51% baseline human: ~ 94%
  - Web: 93.55% vs. 91.82% baseline human: ~ 96%
  - this is *not* just because of the extended tag set!
- Human annotator agreement not satisfactory yet
- Help us further improve the state-of-the-art & data sets!

Now: talks by task winners + poster session + discussion

Colloquial contractions in the Dortmund Chat Corpus

'social chat' subcorpus of the Dortmund chat corpus: 21 logfiles / 104.094 tokens, including 584 occurrences of colloquial contractions



Strukturtyp:	Vorkommen:
VV PPER	304
APPR ART	75
VM PPER	39
VA PPER	36
KOUS PPER	35
PPER PPER	28
ADV ART	17
PTKNEG ADV	9
PWS VV ADV	8
VV ADV	6
VV ART	4
VV PPER ADV	4
PPER ADV	3
PWAV PPER	3
PWS APPR ART	3
VM PPER PPER	2
VV PPER PPER	2
KOKOM ART	1
KOUS ART	1
PWAV VV	1
VA ADV	1
VA PPER PPER	1
VM PPER ART	1
GESAMT:	**584**

91,5% of all cases

Detailed results: Tokenization

	CMC				Web				Overall		
Team	pr	rc	F_1	Rk	pr	rc	F_1	Rk	F_1	Rk	**Pdm**
SoMaJo	99.52	99.56	99.54	1	99.57	99.64	99.60	3	99.57	1	**1**
AIPHES	99.30	98.62	98.96	2	99.63	99.89	99.76	2	99.36	2	**2**
COW	98.31	98.07	98.18	5	99.84	99.71	99.77	1	98.98	3	**3**
WASTE†	99.41	97.57	98.47	4	99.59	99.26	99.42	4	98.95	4	-
LTL-UDE	99.01	98.18	98.58	3	98.92	99.54	99.22	8	98.90	5	**4**
$WAGMOB*	98.97	96.79	97.83	6	99.41	99.38	99.39	5	98.61	6	-
Stanford†	97.19	97.69	97.41	7	98.97	99.71	99.34	7	98.38	7	-
TreeTagger†	94.95	95.01	94.96	8	99.58	99.14	99.36	6	97.16	8	-

Table 7: Results of the tokenization subtask including non-competitive submissions (marked with *) and baseline systems (marked with †). The last column gives the official EmpiriST 2015 "podium" ranking. pr, rc, and F_1 are given as percentages for better readability.

Detailed results: POS tagging

	CMC				Web				Overall		
	STTS_IBK		STTS 1.0		STTS_IBK		STTS 1.0		STTS_IBK		
Team	acc	Rk	acc	Rk	acc	Rk	acc	Rk	acc	Rk	**Pdm**
UdS-distributional$_2$	87.33	1	90.28	1	93.55	1	94.62	1	90.44	1	**1**
UdS-retrain$_2$	86.40	3	89.07	3	92.79	3	93.86	3	89.60	2	-
UdS-surface$_2$	86.45	2	89.28	2	92.43	4	93.50	4	89.44	3	-
LTL-UDE$_2$	86.07	4	88.84	4	92.10	5	93.12	5	89.09	4	**2**
AIPHES	84.22	7	87.10	6	93.27	2	94.30	2	88.75	5	**3**
bot.zen$^*_3$	85.42	5	87.47	5	90.63	8	91.74	9	88.03	6	-
COW†	77.89	8	81.51	8	91.82	6	92.96	6	84.86	7	-
$WAGMOB*	84.77	6	87.03	7	84.51	10	85.57	10	84.64	8	-
TreeTagger†	73.21	9	76.81	9	91.75	7	92.89	7	82.48	9	-
Stanford†	70.60	10	75.83	10	89.42	9	92.52	8	80.01	10	-

Table 8: Results of the PoS tagging subtask including non-competitive or late submissions (marked with *) and baseline systems (marked with †). If applicable, a subscript indicates the best run of the respective system (based on overall accuracy), which is listed in the table. The last column gives the official EmpiriST 2015 "podium" ranking. acc is given as a percentage for better readability.