Processing Social Media Data: Can we Circumvent the Tower of Babel?

Nikola Ljubešić

Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb

Solomon seminar, 10 July 2017

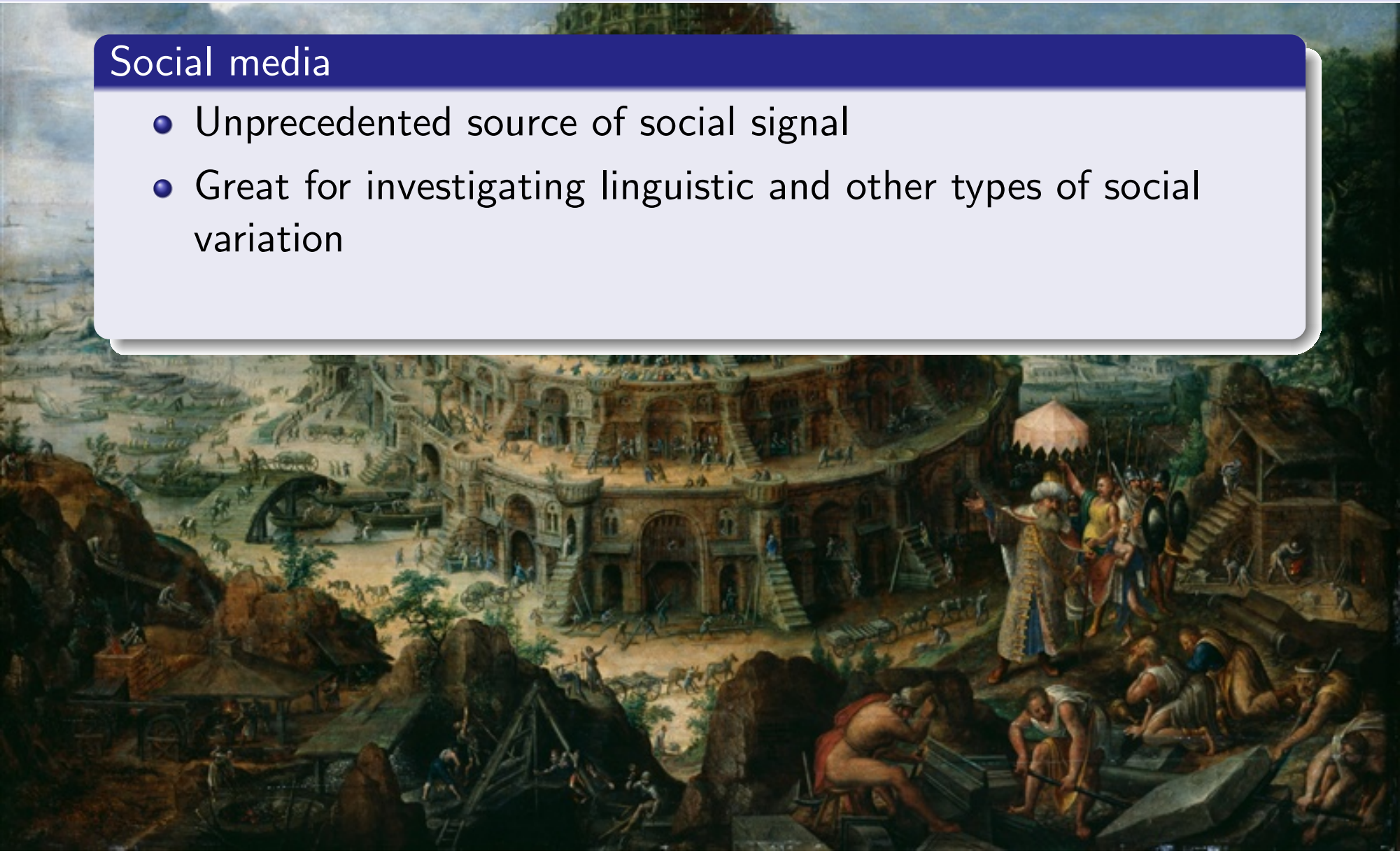
Problem

Problem

Social media

- Unprecedented source of social signal

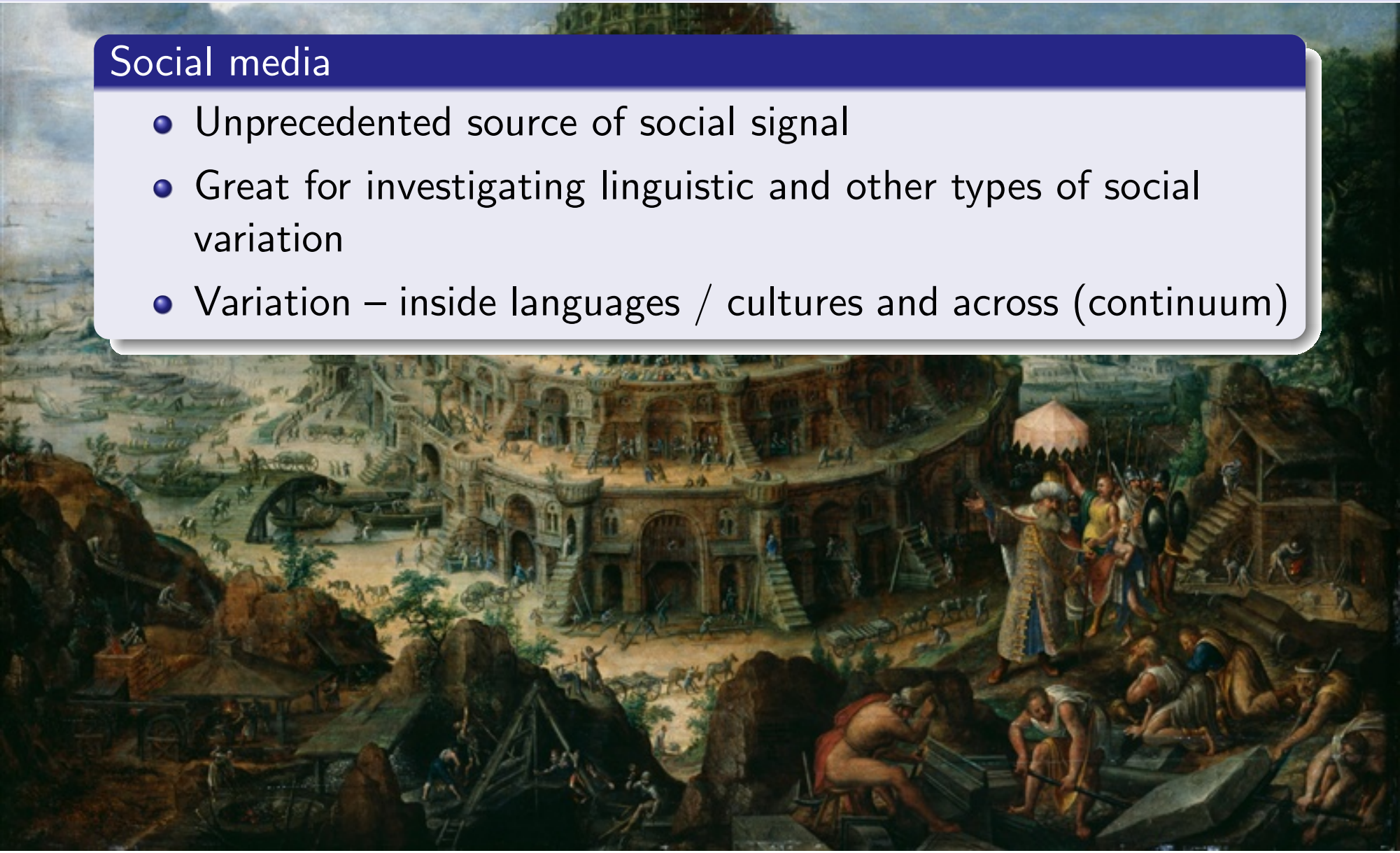
Problem



Social media

- Unprecedented source of social signal
- Great for investigating linguistic and other types of social variation

Problem



Social media

- Unprecedented source of social signal
- Great for investigating linguistic and other types of social variation
- Variation – inside languages / cultures and across (continuum)

Problem

Social media

- Unprecedented source of social signal
- Great for investigating linguistic and other types of social variation
- Variation – inside languages / cultures and across (continuum)

The curse of Babel (for research on social media)

Problem

Social media

- Unprecedented source of social signal
- Great for investigating linguistic and other types of social variation
- Variation – inside languages / cultures and across (continuum)

The curse of Babel (for research on social media)

- Before performing research on a variable, we process data to obtain the maximum amount of explanatory variables
- That same social variation that we are interested in makes the data much harder to process

Problem

Social media

- Unprecedented source of social signal
- Great for investigating linguistic and other types of social variation
- Variation – inside languages / cultures and across (continuum)

The curse of Babel (for research on social media)

- Before performing research on a variable, we process data to obtain the maximum amount of explanatory variables
- That same social variation that we are interested in makes the data much harder to process
- Can we circumvent this curse by performing model adaptations and / or model on stable predictors?

Overview

Overview

Linguistic processing

- Part-of-speech tagging, parsing, named entity linking etc.
- Accuracy of English PoS-tagging – WSJ 97%, Twitter 85% (Gimpel et al. 2011)
- How can we adapt to such extremely different domains?
- Is there a path towards not only domain adaptation, but truly cross-lingual and cross-domain models?

Overview

Linguistic processing

- Part-of-speech tagging, parsing, named entity linking etc.
- Accuracy of English PoS-tagging – WSJ 97%, Twitter 85% (Gimpel et al. 2011)
- How can we adapt to such extremely different domains?
- Is there a path towards not only domain adaptation, but truly cross-lingual and cross-domain models?

User profiling

- Enriching user information a constant requirement
- Traditionally relies on linguistic content communicated
- Can we build predictive models that work on different languages / cultures, relying on (more) stable predictors?

Very brief history of text processing

Knowledge-driven systems

- (Mostly) hand-crafted knowledge
- Hard to adapt to different domains, especially languages
- Dominant paradigm until the 90s

Very brief history of text processing

Knowledge-driven systems

- (Mostly) hand-crafted knowledge
- Hard to adapt to different domains, especially languages
- Dominant paradigm until the 90s

Data-driven systems

- Instead of writing rules, build a statistical model from data
- Adapting to different domains or languages "only" requires target data

Very brief history of text processing

Knowledge-driven systems

- (Mostly) hand-crafted knowledge
- Hard to adapt to different domains, especially languages
- Dominant paradigm until the 90s

Data-driven systems

- Instead of writing rules, build a statistical model from data
- Adapting to different domains or languages "only" requires target data
- Two (three?) phases
  - WSJ (Wall Street Journal) era – narrow domain datasets used both for modeling and evaluation
  - Social media era – necessity of domain adaptation
  - Cross-lingual era – neural models

How to approach processing non-standard text

How to approach processing non-standard text

Normalisation

- Transformation of non-standard input to a standard one
- Further processing with the standard ("WSJ") pipeline

How to approach processing non-standard text

Normalisation

- Transformation of non-standard input to a standard one
- Further processing with the standard ("WSJ") pipeline

Direct processing

- Domain adaptation
- Supervised – in-domain manually tagged data
- Unsupervised – distributional information (lexis), unambiguous sentences (syntax) etc.

How to approach processing non-standard text

Normalisation

- Transformation of non-standard input to a standard one
- Further processing with the standard ("WSJ") pipeline

Direct processing

- Domain adaptation
- Supervised – in-domain manually tagged data
- Unsupervised – distributional information (lexis), unambiguous sentences (syntax) etc.

Joint processing

- Performing normalisation and tagging simultaneously – dependent processes

Normalisation

- Normalisation as a translation task, just generalised to the character level, statistical approach (Moses)

  `_ j s t _ n e v e m _ -> _ j a z _ n e _ v e m _`
- `https://github.com/clarinsi/csmtiser`
- Ljubešić et al. 2016; Scherrer and Ljubešić, 2016
- Gain from multiple LMs, even distant (different dialects)
- Sentence-level when data strongly deviates from the norm

dataset	none	baseline	CSMT	+LMs	error reduction	
18th century	17.63	6.46	1.55	1.33	93%	79%
19th century	3.13	1.43	1.01	0.91	71%	36%
Twitter L3	5.15	2.44	2.19	1.65	70%	32%
Twitter L1	0.75	0.62	0.41	0.34	55%	45%

Normalisation

- Normalisation as a translation task, just generalised to the character level, statistical approach (Moses)

  `_ j s t _ n e v e m _ -> _ j a z _ n e _ v e m _`
- `https://github.com/clarinsi/csmtiser`
- Ljubešić et al. 2016; Scherrer and Ljubešić, 2016
- Gain from multiple LMs, even distant (different dialects)
- Sentence-level when data strongly deviates from the norm

dataset	none	baseline	CSMT	+LMs	error reduction	
18th century	17.63	6.46	1.55	1.33	93%	79%
19th century	3.13	1.43	1.01	0.91	71%	36%
Twitter L3	5.15	2.44	2.19	1.65	70%	32%
Twitter L1	0.75	0.62	0.41	0.34	55%	45%

1st place in CLIN27 shared task on translating historical Dutch, outperforming NMT models

Tagging

- Ljubešić et al. 2017 vs. Plank et al. 2016 on Slovene Twitter
- CRF vs. BI-LSTM with word and character embeddings
- Domain adaptation through distributional information – Brown clusters (jaz js jst jes jsss), word embeddings

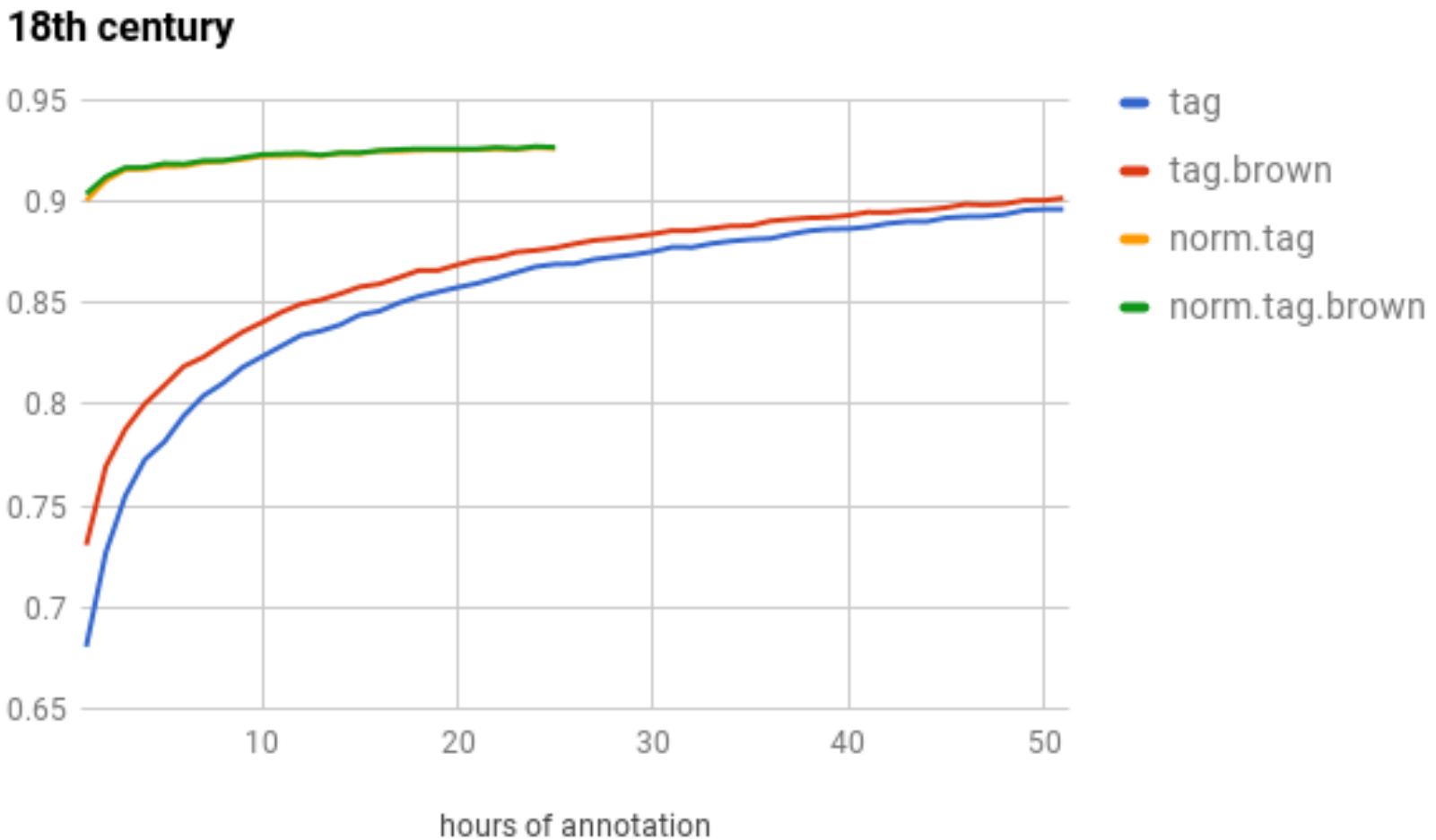
	Ljubešić et al. 2017		Plank et al. 2016	
Configuration	**MSD**	**PoS**	**MSD**	**PoS**
standard test data	94.27	98.94	92.92	97.80
+distrbutional			94.29	98.02
non-standard test data	68.67	73.13	67.06	73.47
+distributional			71.27	80.46
in-domain training data	84.15	89.85		
+distributional	85.70	91.52	86.03	92.12
+normalisation	86.28	91.72		
+standard	**87.70**	**92.22**	**88.15**	**92.12**

Normalisation data vs. tagging data (work in progress)

- Task of PoS tagging non-standard text, should we invest into producing in-domain normalisation or tagging datasets?
- Hypothesis: limited resources normalisation, otherwise tagging

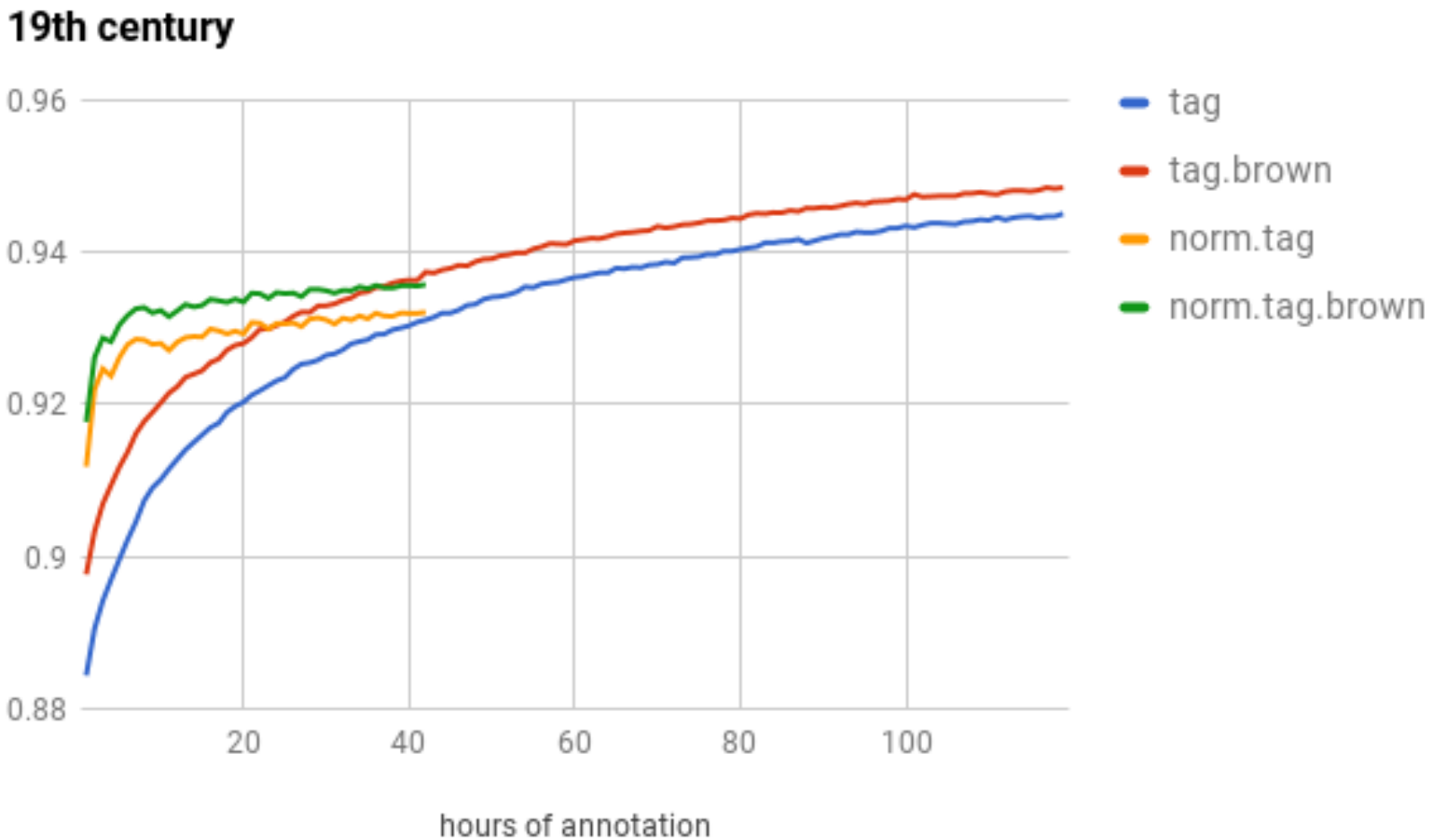
Normalisation data vs. tagging data (work in progress)

- Task of PoS tagging non-standard text, should we invest into producing in-domain normalisation or tagging datasets?
- Hypothesis: limited resources normalisation, otherwise tagging



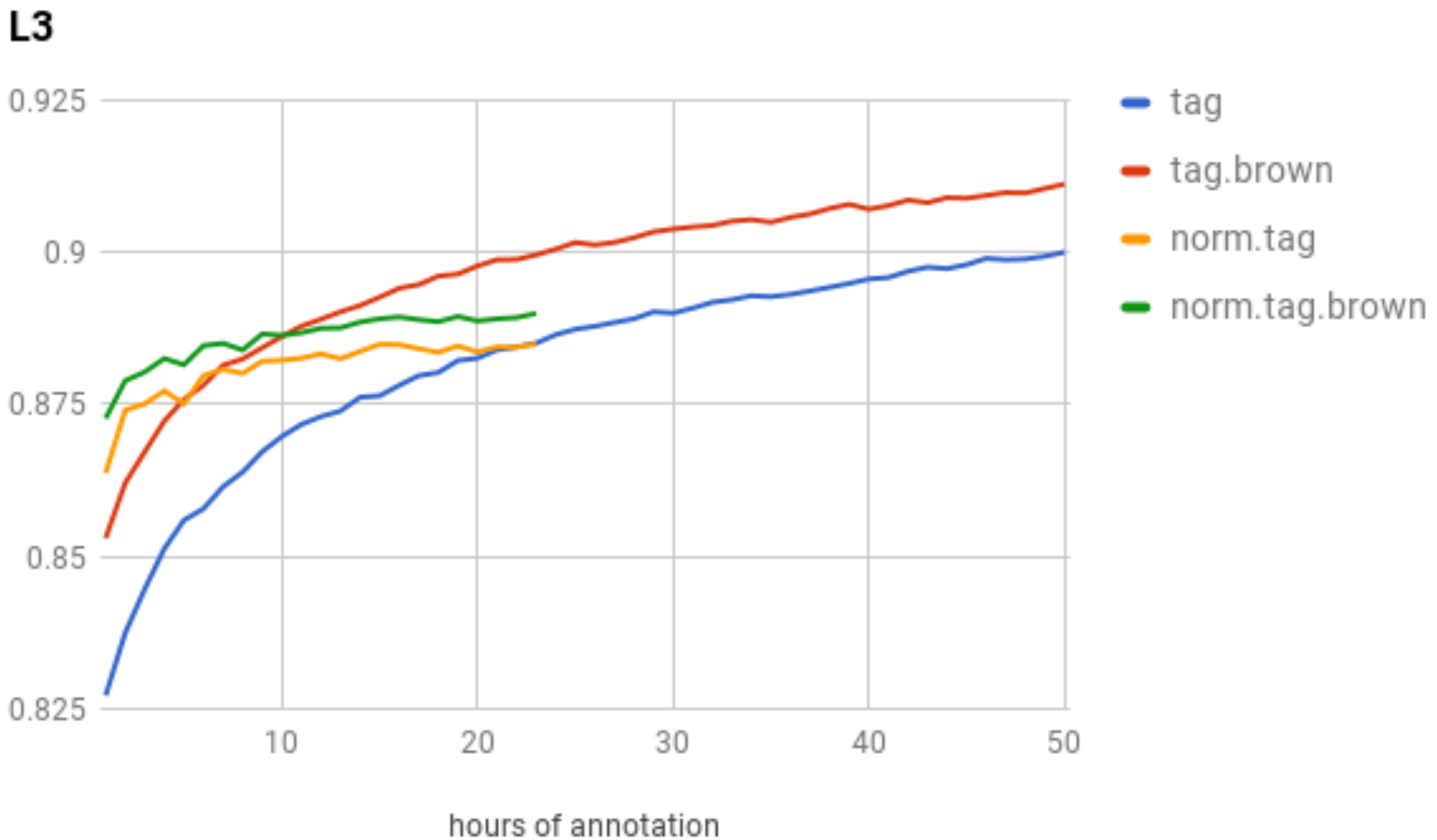
Normalisation data vs. tagging data (work in progress)

- Task of PoS tagging non-standard text, should we invest into producing in-domain normalisation or tagging datasets?
- Hypothesis: limited resources normalisation, otherwise tagging



Normalisation data vs. tagging data (work in progress)

- Task of PoS tagging non-standard text, should we invest into producing in-domain normalisation or tagging datasets?
- Hypothesis: limited resources normalisation, otherwise tagging



Cross-lingual processing

- Covered minor (in-language) variation, what about major (cross-language) one?
- Gillick et al. (2015) read text byte-by-byte and output spans with annotations, train single model for multiple languages, positive interference across languages
- Large body of work on multilingual word representations, Bojanowski et al. (2016) build representations for 294 languages, Smith et al. (2017) transform these representations to maximise lexical similarity for 78 languages

Cross-lingual processing

- Covered minor (in-language) variation, what about major (cross-language) one?
- Gillick et al. (2015) read text byte-by-byte and output spans with annotations, train single model for multiple languages, positive interference across languages
- Large body of work on multilingual word representations, Bojanowski et al. (2016) build representations for 294 languages, Smith et al. (2017) transform these representations to maximise lexical similarity for 78 languages
- Use case: word abstractness–concreteness prediction
- Training data: 3000 words in Croatian, Likert scale 1-5, use word embeddings as features, Spearman correlation 0.81
- Apply the model learned on the transformed Croatian representation to remaining 77 languages

Cross-lingual processing

pristnost	unquestionable	fragwürdig
nepristranskost	intentionality	voraussetzen
avtentičnost	unquestioned	gerechtfertigt
primernost	definitional	erstrebenswert
koristnost	arbitrariness	voraussetzt
psevdoznanost	subjectivity	erachten
pravičnost	actualization	vorstellbar
miselnost	inevitability	rechtfertigen
intuicija	rationality	ausschlaggebende
pristranskost	conceptions	bestimmend
...	...	...
slamica	cans	beutel
podstrešju	ashtray	schläger
škatlice	buckle	ballen
torbica	bottle	henne
predpasnik	sidewalk	deckel
škatlica	jacket	teller

User profiling experiments

User profiling experiments

User type identification (Ljubešić and Fišer, 2016)

- Apply a simple {private, corporate} schema
- Slovene (and Croatian) Twitter users
- BoW vs. language-agnostic features
- Experiments on time (training before, testing after 2015) and space robustness (training on Slovene, testing on Croatian)

User profiling experiments

User type identification (Ljubešić and Fišer, 2016)

- Apply a simple {private, corporate} schema
- Slovene (and Croatian) Twitter users
- BoW vs. language-agnostic features
- Experiments on time (training before, testing after 2015) and space robustness (training on Slovene, testing on Croatian)

Gender prediction (Ljubešić et al. 2017)

- TwiSty corpus – German, Italian, Dutch, French, Spanish, Portuguese Twitter users
- BoW vs. language-agnostic features
- Experiments on cross-lingual gender prediction

Language-agnostic features

- `perc`
  - `perc_http` – percentage of tweets containing URLs
  - `perc_reply` – percentage of tweets being replies
- `mean`
  - `mean_hour` – mean of the posting hour
  - `mean_len_text` – mean of the length of the tweet
- `med` – median, same as `mean`
- `var` – variance, same as `mean`
- `user`
  - `user_ff_ratio` – ratio of friends and followers
  - `user_red_back` – intensity of red color component in user's background

User type identification

Feature type	Corporate	Private	Both
MFC baseline	0.0000	0.8572	0.6430
language agnostic (LA)	0.7944	0.9335	0.8987
BoW	0.8742	0.9577	0.9368
ensemble	0.8864	0.9620	**0.9431**

User type identification

Feature type	Corporate	Private	Both
MFC baseline	0.0000	0.8572	0.6430
language agnostic (LA)	0.7944	0.9335	0.8987
BoW	0.8742	0.9577	0.9368
ensemble	0.8864	0.9620	**0.9431**

		Time dependence		Space dependence	
Feature type	All	$\geq$ 2015	Loss	Croatian	Loss
MFC	0.6430	0.6430	-	0.3388	-
LA	0.8987	0.8370	7%	0.6270	30%
BoW	0.9368	0.9280	1%	0.7304	22%

Feature analysis

Private users...

- Reply more
- Mention other users
- Favour other tweets
- Post in various hours
- Post tweets of various lengths

Feature analysis

Private users...

- Reply more
- Mention other users
- Favour other tweets
- Post in various hours
- Post tweets of various lengths

Corporate users...

- Use more URLs
- Post during working hours
- Post earlier in the day
- Post longer tweets

Gender prediction

Lang	Inst. #	MFC	ILBoW	CLBoW
DE	376	36.63	77.91	61.26
IT	429	50.96	62.46	58.66
NL	933	34.59	80.68	61.55
FR	1207	41.78	78.70	56.61
PT	3572	43.97	85.26	53.18
ES	9639	41.13	83.04	57.99

Gender prediction

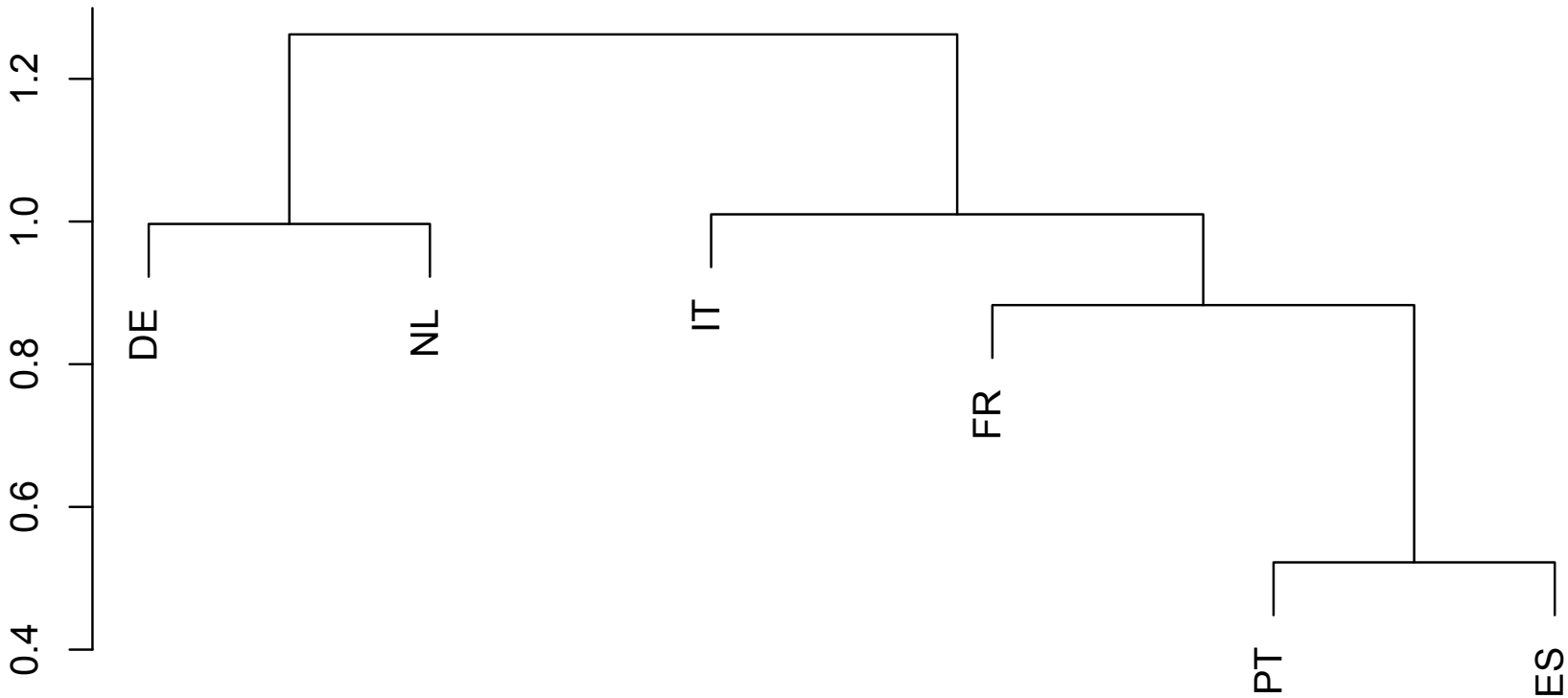
Lang	Inst. #	MFC	ILBoW	CLBoW
DE	376	36.63	77.91	61.26
IT	429	50.96	62.46	58.66
NL	933	34.59	80.68	61.55
FR	1207	41.78	78.70	56.61
PT	3572	43.97	85.26	53.18
ES	9639	41.13	83.04	57.99

Lang	DE	IT	NL	FR	PT	ES
DE	69.37	**63.30**	**67.26**	**68.35**	65.59	**69.92**
IT	**66.98**	**63.91**	**66.76**	**63.73**	**63.47**	**66.12**
NL	62.10	**61.15**	68.02	**57.87**	59.64	**64.68**
FR	**69.70**	**65.12**	62.68	67.47	**65.60**	**66.35**
PT	**61.94**	**57.31**	57.23	**62.65**	69.51	68.12
ES	**62.89**	55.80	**64.85**	**66.82**	67.27	71.47

Feature effect sizes

Feature	DE	IT	NL	FR	PT	ES
perc_emoji	0.63	0.21	0.45	0.49	0.41	0.5
mean_retweet_count	0.09	0.03	0.09	0.38	0.27	0.22
user_red_back	0.24	0.09	0.13	0.23	0.38	0.42
perc_http	-0.21	-0.24	-0.25	-0.15	-0.27	-0.17
perc_ios	-0.23	-0.22	-0.09	-0.19	-0.09	-0.13
var_retweet_count	-0.1	0.05	0.1	0.11	0.03	0.04
perc_retweeted	-0.01	0.2	-0.2	0.2	0.26	0.17
perc_question	-0.35	-0.13	-0.1	-0.29	-0.14	-0.11
user_tweet_per_day	0.08	0.19	0.01	0.31	0.15	0.12
perc_emoticon	-0.23	-0.25	-0.17	-0.18	-0.24	-0.1
user_location	-0.17	-0.2	-0.21	-0.11	-0.17	-0.12
mean_hour	0.08	0.23	0.18	0.22	-0.1	-0.02
var_len_text	0.25	0.24	0.2	0.24	0.01	0.08

Clustering countries by feature effect sizes



Conclusion

Linguistic processing

- Minor variation (inside language (family)) can partially be dealt with via unsupervised adaptation
- Major variation (across languages) still requires supervised
- Promising directions for major variation:
  - Multilingual learning – positive interference
  - Multilingual representations – successful in simpler tasks

User profiling

- Minor variation – BoW stronger than LA features
- Major variation – LA features outperform BoW
- Feature adaptation / generalisation
- Multidimensional and multimodal signal

Acknowledgement

Špela Arhar Holdt, Jaka Čibej, Tomaž Erjavec, **Darja Fišer**, Barbara Plank, Tanja Samardžić, Yves Scherrer, Katja Zupan

Acknowledgement

CLARIN.SI

Next venture

- Interdisciplinary research of socially unacceptable discourse (broader than hate speech)
- Technological, sociological, linguistic, legal perspective

Processing Social Media Data: Can we Circumvent the Tower of Babel?

Nikola Ljubešić

Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb

Solomon seminar, 10 July 2017