Natural Language Processing challenges

in the context of Polish and other Slavic languages

Łukasz Kobyliński, Michał Wasiluk

About us

Łukasz (@lkobylinski)

- Research @ IPI PAN (ICS PAS)
- Consulting @ SigDelta
- Management @ Sages
- Teaching @ PW (WUT)
- #NLP, #ML, #Education

Michał

- Research @ IPI PAN (ICS PAS), PW (WUT), SGH (WSoE)
- #ML, #DL, #Programming

NLP is important

it is no longer academia any more

- speech recognition
- question answering

NLP is important

it is no longer academia any more

- speech recognition
- question answering
- sentiment analysis
- named entity recognition
- coreference resolution

and more...

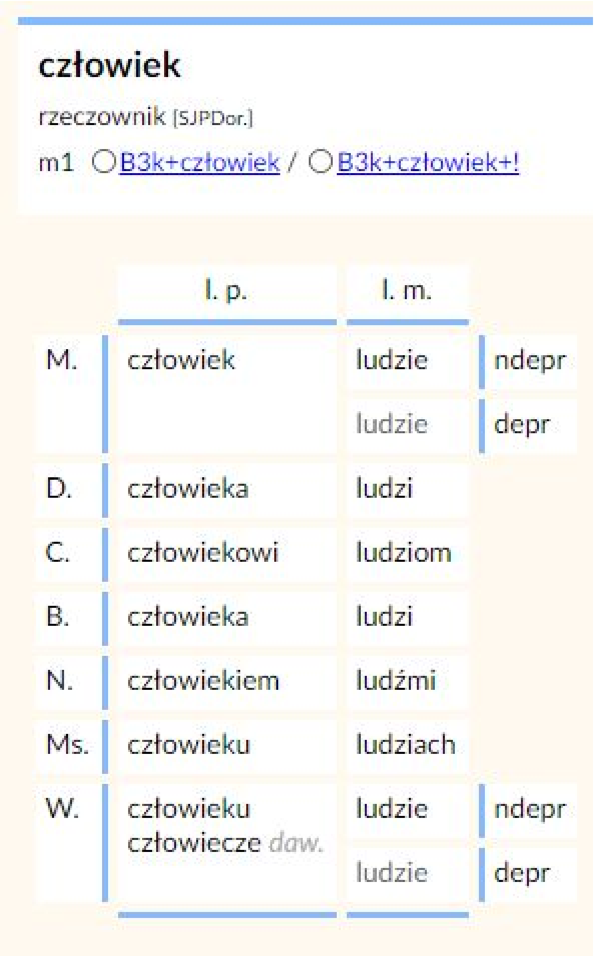
- event identification
- sense disambiguation
- information extraction

Polish (and other Slavic languages) are hard!

They are highly inflected

… and, as a consequence:

- the number of possible word forms in a corpus is much higher than e.g. English,
- many word forms are morphologically ambiguous
  - człowieka (D. - genitive / B. - accusative)
  - człowieku (Ms. - locative / W. - vocative)
  - ludzie (M. - nominative / W. pl. - vocative plural)
  - ludzi (D. - genitive / B. pl. - accusative plural)



człowiek

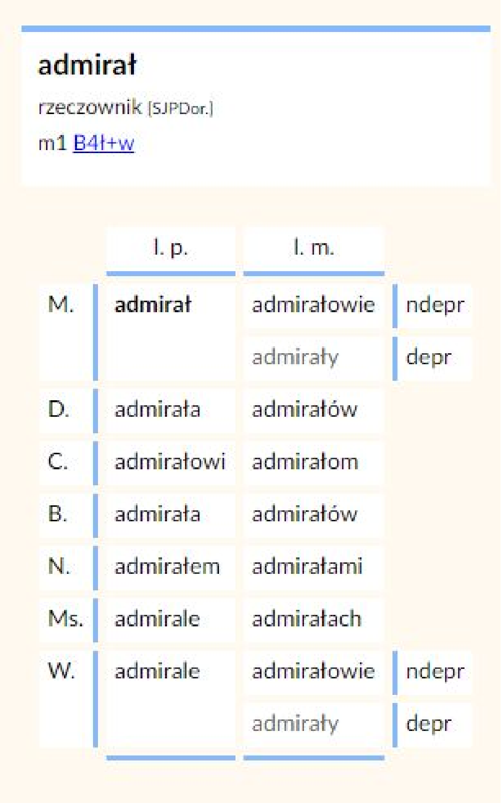
rzeczownik [SJPDor.]

m1 ○B3k+człowiek / ○B3k+człowiek+!

	l. p.	l. m.	
M.	człowiek	ludzie	ndepr
		ludzie	depr
D.	człowieka	ludzi	
C.	człowiekowi	ludziom	
B.	człowieka	ludzi	
N.	człowiekiem	ludźmi	
Ms.	człowieku	ludziach	
W.	człowieku człowiecze *daw.*	ludzie	ndepr
		ludzie	depr

Polish (and other Slavic languages) are hard!

Proper word homonymy is an additional problem

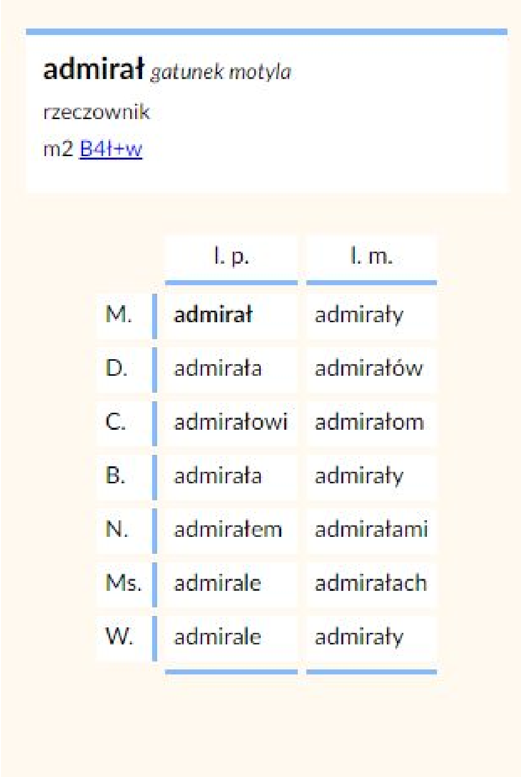


admirał

rzeczownik [SJPDor.]

m1 B4ł+w

	l. p.	l. m.	
M.	**admirał**	admirałowie	ndepr
		admirały	depr
D.	admirała	admirałów	
C.	admirałowi	admirałom	
B.	admirała	admirałów	
N.	admirałem	admirałami	
Ms.	admirale	admirałach	
W.	admirale	admirałowie	ndepr
		admirały	depr



admirał *gatunek motyla*

rzeczownik

m2 B4ł+w

	l. p.	l. m.
M.	**admirał**	admirały
D.	admirała	admirałów
C.	admirałowi	admirałom
B.	admirała	admirały
N.	admirałem	admirałami
Ms.	admirale	admirałach
W.	admirale	admirały

Polish (and other Slavic languages) are hard!

Free word order!

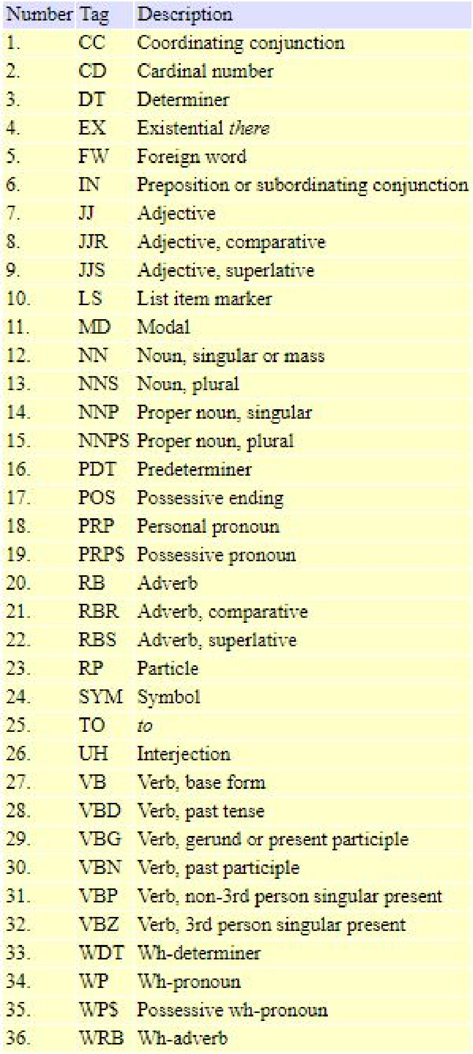
- English
  - John loves Mary (SVO)
- Polish
  - Jan kocha Marię (SVO),
  - Jan Marię kocha,
  - Marię kocha Jan,
  - Marię Jan kocha,
  - Kocha Marię Jan,
  - Kocha Jan Marię.

Fixed-context systems (like HMMs) are not as effective as for English.

Example: Part of speech tagging

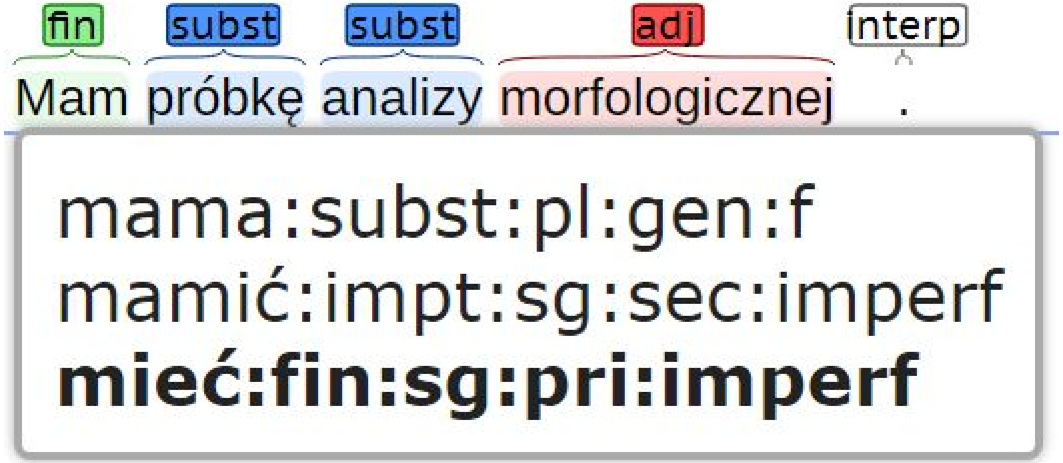
A task of marking up words in text with a corresponding part of speech

- not as simple as `dog = noun`
- a definition of tagset is needed
- in English: 36 - 200 tags, e.g. the Penn Treebank:
  - NN - noun (singular); NNS - noun (plural);
  - VB - verb, base form; VBD - verb, past tense
- in Polish: 4 000 theoretically possible tags!
- (similar in Czech, ca. 2 000 in Slovene).



Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

POS in Polish

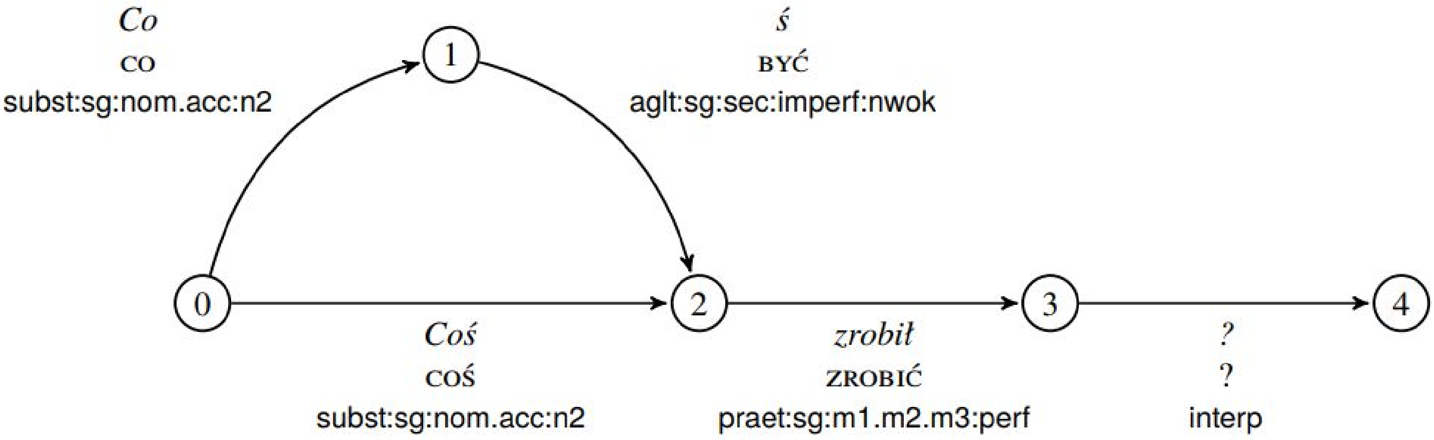


mama:subst:pl:gen:f
mamić:impt:sg:sec:imperf
mieć:fin:sg:pri:imperf

The tagset is positional

- `mam – mieć:`**`fin:sg:pri:imperf`**
  - `mieć –` lemma
  - `fin –` grammatical class (35 possibilities)
  - `sg, pri, imperf –` required grammatical attributes (specific for each class)
  - for many grammatical classes - also optional attributes
- … that leads to ca. 4 000 possible tag values
  (ca. 1 000 appear in a real-world corpus).

Bonus - segmentation ambiguities!



Marcin Woliński, Morfeusz Reloaded (2014).

What (co) have you done (-ś zrobił)?’ or ‘Did he do (zrobił) anything (coś)?’.

Accuracy of POS Tagging

In English: exceeding 97%

In Polish

2007: Rule-based tagger (TaKiPI), ca. **88%** tagging accuracy

2010: Brill tagger adapted to Polish (PANTERA), ca. **89%** tagging accuracy

2012: Classic Machine Learning (memory based, CRF - WMBT, WCRFT, Concraft), ca. **91%** tagging accuracy

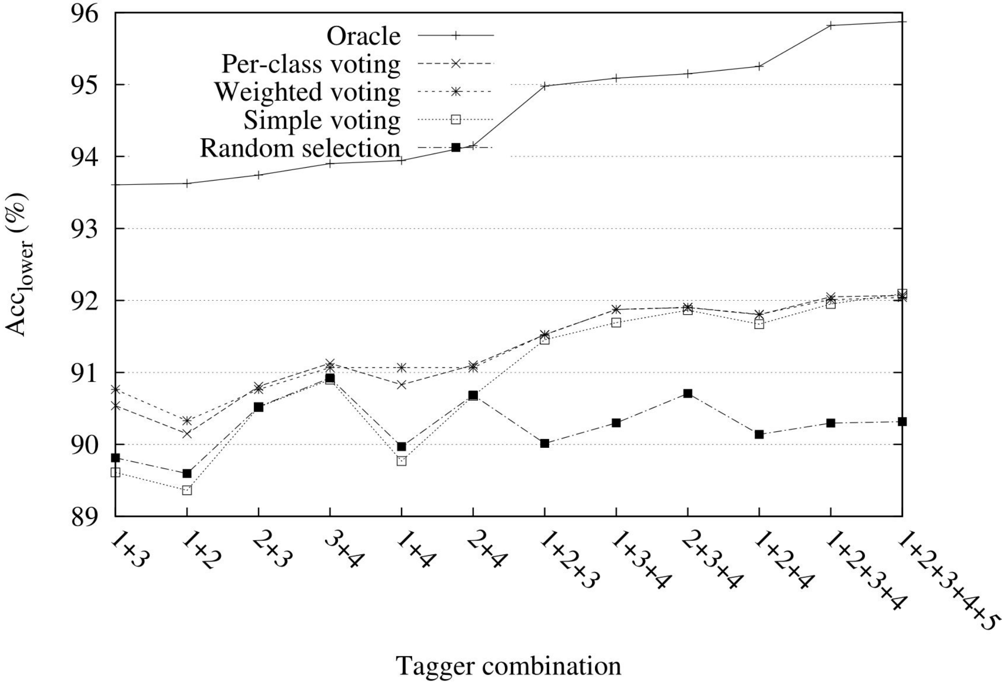
… not enough.

Accuracy of POS Tagging in Polish

2014: Ensembling (PoliTa), ca. **92%** tagging accuracy

… still not enough :)

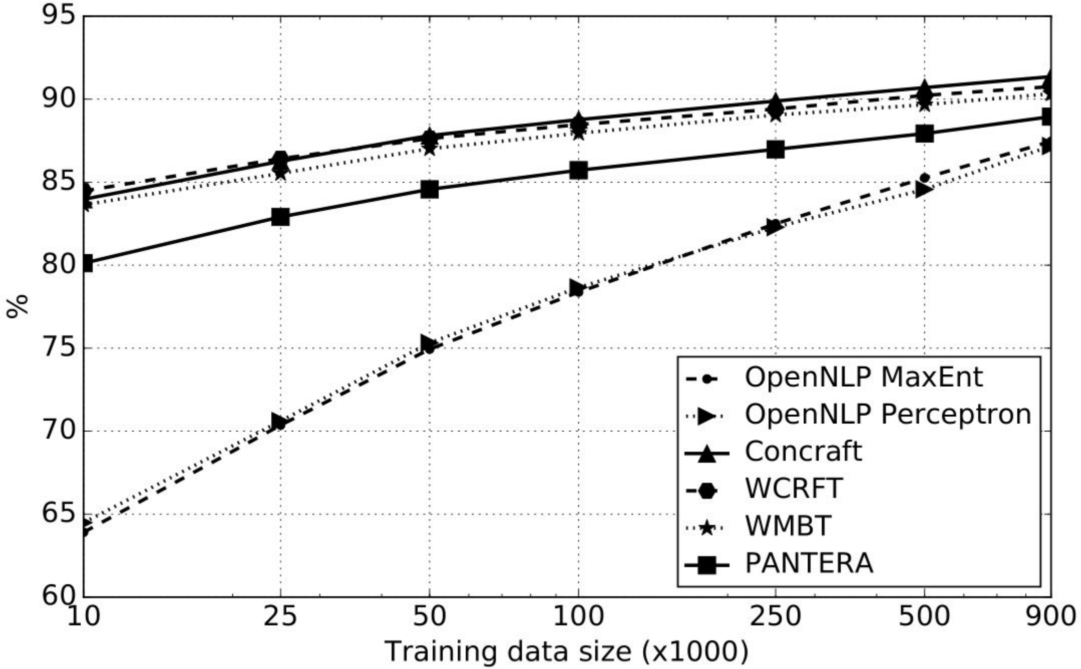
8% error rate translates to 80 million mistakes in a 1 billion corpus!



Lack of resources

The size of the (annotated) gold-standard training data matters a lot.

… but it is very expensive to produce.



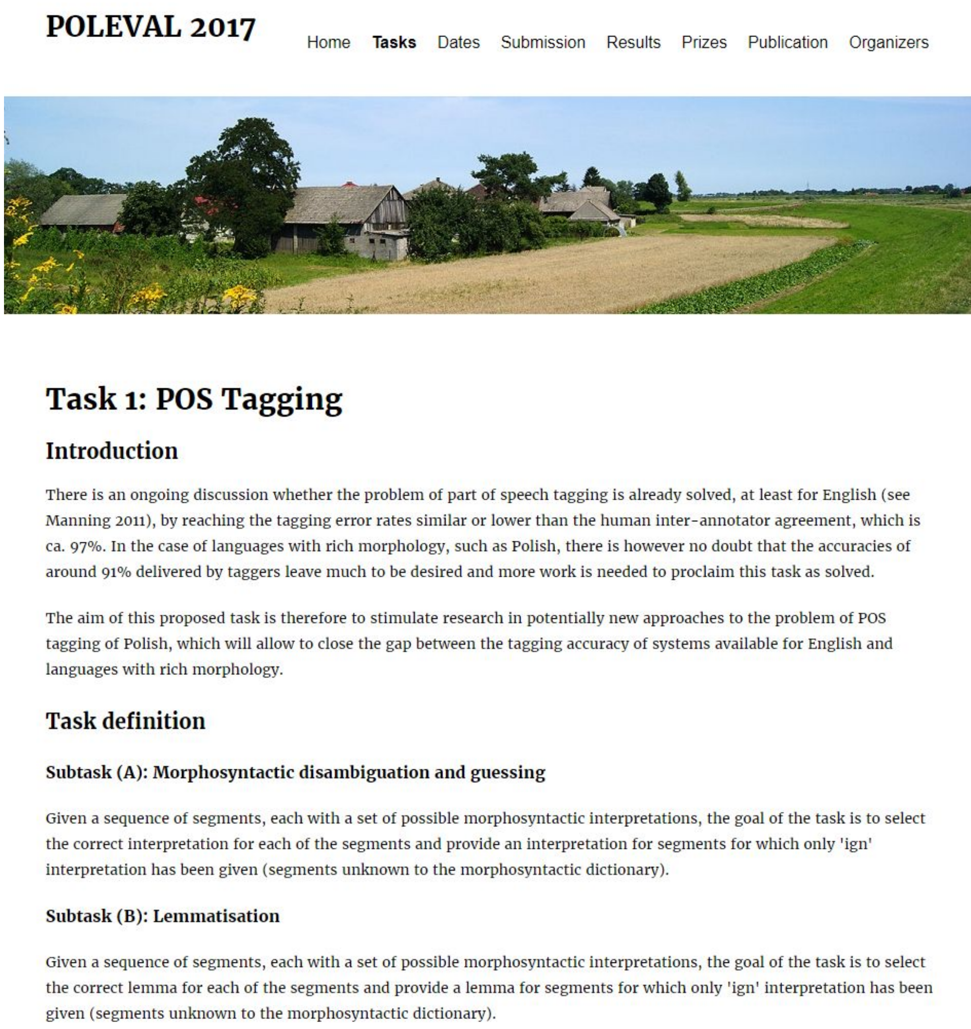
PolEval 2017

Part-of-speech tagging and sentiment analysis shared tasks.

… basically a Kaggle-like competition.

Submissions:

- 16 submissions from 9 teams,
- all submissions were based on neural networks!



Task 1: POS Tagging

Introduction

There is an ongoing discussion whether the problem of part of speech tagging is already solved, at least for English (see Manning 2011), by reaching the tagging error rates similar or lower than the human inter-annotator agreement, which is ca. 97%. In the case of languages with rich morphology, such as Polish, there is however no doubt that the accuracies of around 91% delivered by taggers leave much to be desired and more work is needed to proclaim this task as solved.

The aim of this proposed task is therefore to stimulate research in potentially new approaches to the problem of POS tagging of Polish, which will allow to close the gap between the tagging accuracy of systems available for English and languages with rich morphology.

Task definition

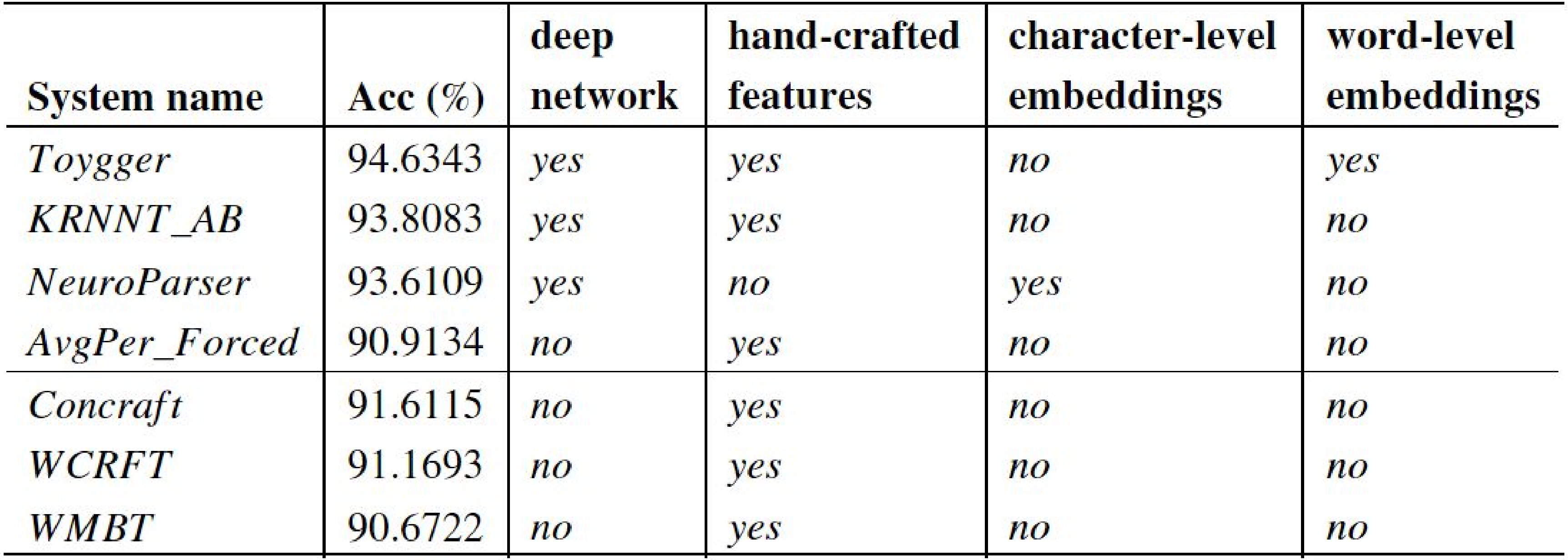
Subtask (A): Morphosyntactic disambiguation and guessing

Given a sequence of segments, each with a set of possible morphosyntactic interpretations, the goal of the task is to select the correct interpretation for each of the segments and provide an interpretation for segments for which only 'ign' interpretation has been given (segments unknown to the morphosyntactic dictionary).

Subtask (B): Lemmatisation

Given a sequence of segments, each with a set of possible morphosyntactic interpretations, the goal of the task is to select the correct lemma for each of the segments and provide a lemma for segments for which only 'ign' interpretation has been given (segments unknown to the morphosyntactic dictionary).

PolEval 2017 POS results



System name	Acc (%)	deep network	hand-crafted features	character-level embeddings	word-level embeddings
Toygger	94.6343	*yes*	*yes*	*no*	*yes*
KRNNT_AB	93.8083	*yes*	*yes*	*no*	*no*
NeuroParser	93.6109	*yes*	*no*	*yes*	*no*
AvgPer_Forced	90.9134	*no*	*yes*	*no*	*no*
Concraft	91.6115	*no*	*yes*	*no*	*no*
WCRFT	91.1693	*no*	*yes*	*no*	*no*
WMBT	90.6722	*no*	*yes*	*no*	*no*

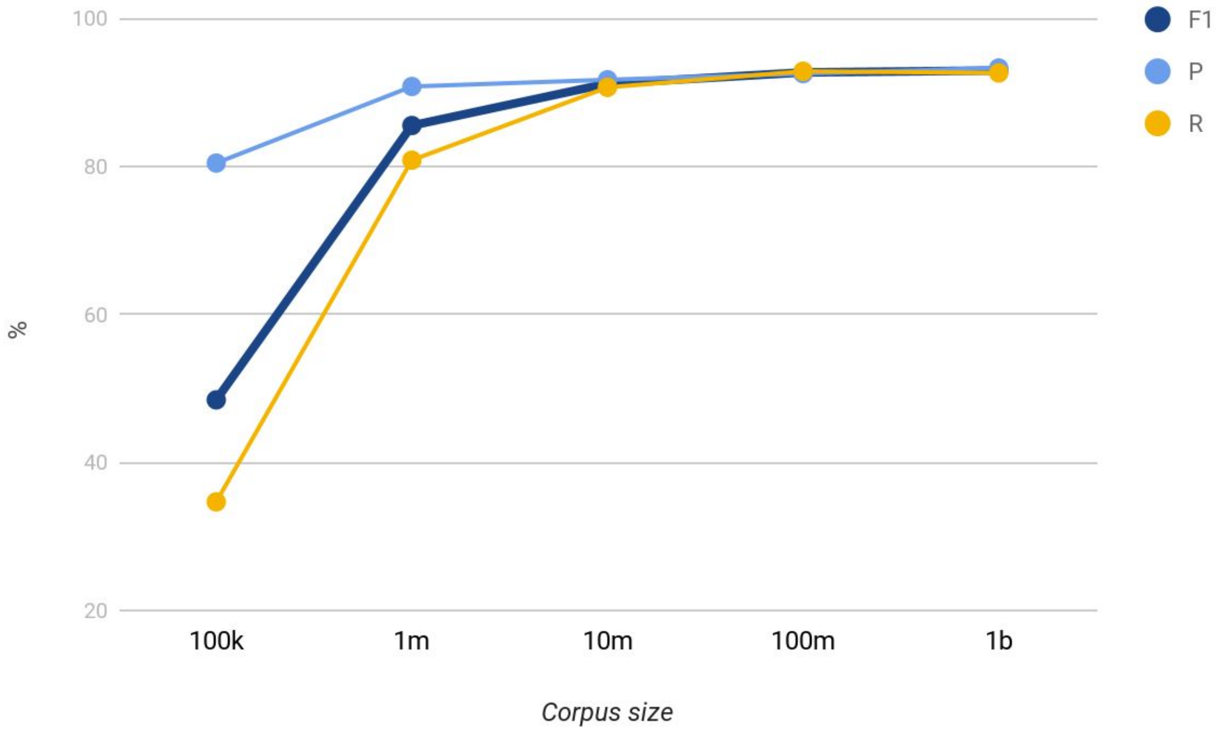
Deep Learning challenges

Word embeddings seem to be the key, but:

- a lot of training data is needed,
- which embeddings are best for Slavic languages? (word2vec, fasttext, …)
- what parameters should be used?
- how to calculate and use the embeddings?
  - orthographic forms of words,
  - word lemmas,
  - grammatical categories / whole POS tags,
  - embeddings based on features (suffix/prefix, synonym/hypernym, ...),
  - combinations of the above.

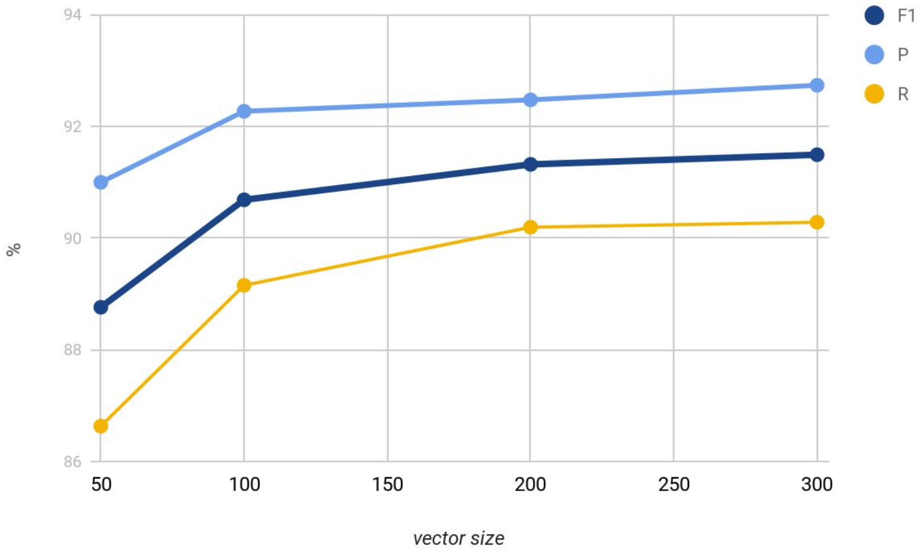
Word embeddings: how large should my training corpus be?

Task: event identification

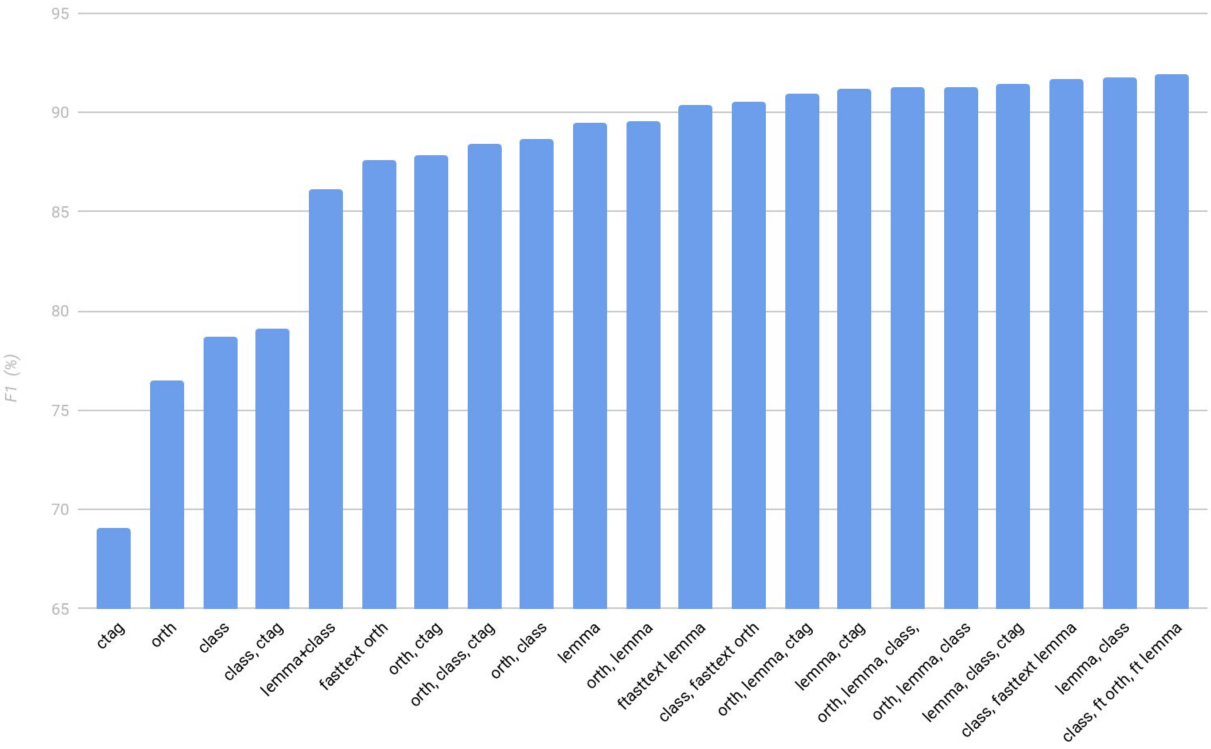


Word embeddings: influence of the vector size

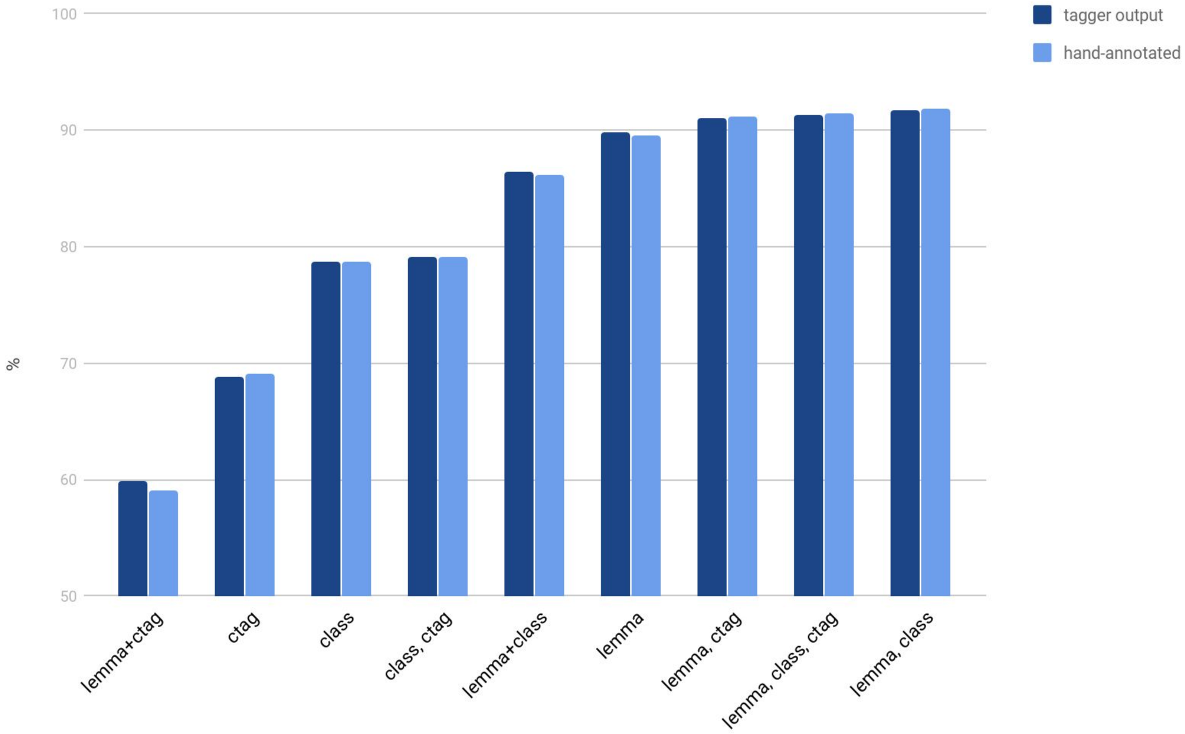
Task: event identification



How to calculate the embeddings?



Is the original annotation quality important?



Conclusions

Slavic languages pose a more difficult problem in processing than English

- size of training data (for your task and for word embeddings) is king,
- deep learning and word embeddings FTW,
- hand-crafted features work well alongside word embeddings,
  - domain knowledge is still important!
- not only word-form based embeddings,
  - lemmas, pos tags, other features
- try different embedding variants besides word2vec,
  - GloVe, fasttext, ...
- you need a lot of tuning and hyperparameter optimization.

How do I start on my own?

Deep learning - Keras (https://keras.io/), also Tensorflow, PyTorch

Word embeddings - Gensim (https://radimrehurek.com/gensim/)

Pre-built models (for Polish) - http://dsmodels.nlp.ipipan.waw.pl/

Tagger implementations - PolEval (http://poleval.pl/)

Other resources - http://clip.ipipan.waw.pl/

How do I start on my own?

Text resources

- Annotated corpora
  - National Corpus of Polish (http://clip.ipipan.waw.pl/NationalCorpusOfPolish)
  - Polish Corpus of the 1960s (http://clip.ipipan.waw.pl/PL196x)
  - PolEval 2017 data (http://clip.ipipan.waw.pl/PolEval)
- Large text collections
  - Wikipedia
  - CommonCrawl

Thank you!