Improving Polish Mention Detection with Valency Dictionary

Bartłomiej Nitoń and Maciej Ogrodniczuk

CORBON 2017 Valencia, Spain, 4th April 2017

The case of mention borders

A mention – text fragment which could potentially create references to discourse world objects.

Inclusion of extensive syntactically dependent phrases into mention borders is important due to semantic understanding of mentions:

- *pierwszy człowiek na Księżycu* ’the first man on the Moon’
- *samochód, który potrącił moją żonę* ’the car which hit my wife’

Mention components (highlights)

- nouns in genitive, e.g. *kolega brata* ‘a friend of my brother’
- adjectives / adjective participles adjusting their form to the superordinate noun, e.g. *kolorowe kwiaty* ‘colourful flowers’, *nadchodzące zmiany* ‘oncoming changes’
- adverbs as adjectives and participle modifiers, e.g. *szalenie ciekawy film* ‘incredibly interesting film’
- prepositional-nominal phrases, e.g. *ustawa o podatku dochodowym* ‘the law on income tax’
- relative clauses, e.g. *dziewczyna, o której rozmawialiśmy* ‘the girl we talked about’

State-of-the-art for Polish

No (sufficiently effective) constituency parser to detect mentions.

Rule based tool combining information on:

- single-segment nouns and nominal groups, detected with Spejd shallow parser fitted with an adaptation of the National Corpus of Polish grammar
- pronouns, identified with a disambiguating morphosyntactic tagger with a morphological analyser and lemmatizer Morfeusz
- zero subjects, detected using machine learned model
- nominal named entities, detected with Nerf named entity recognizer

Mention detection improvements

Observation: valence schemata can bring improvements to mention detection.

- verbal schemata: *confuse sb with sb*
  → never link (sb with sb)
- nominal schemata: *conflict of sb with sb*
  → always link (conflict of sb with sb)

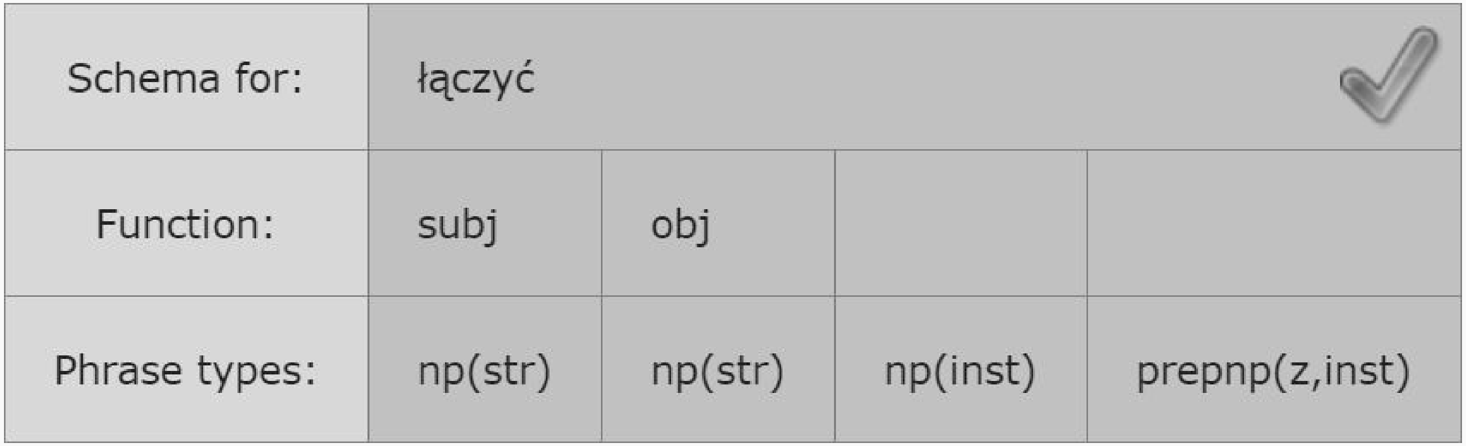
Walenty: a source of syntactic schemata

Walenty is a comprehensive human- and machine-readable dictionary of Polish valency information for verbs, nouns, adjectives and adverbs:

- **over 12 000 verbs (> 67 000 syntactic schemata)**
- **about 3 000 nouns (> 18 000 syntactic schemata)**
- about 1 000 adjectives (> 4 000 syntactic schemata)
- about 200 adverbs (> 1 000 syntactic schemata)

And is still expanding...

Walenty (example schema)



Schema for:	łączyć ✓			
Function:	subj	obj		
Phrase types:	np(str)	np(str)	np(inst)	prepnp(z,inst)

Potężne [komputery]$_{\text{SUBJ}}$ [łączą]$_{\text{VERB}}$ [firmę]$_{\text{OBJ}}$ [światłowodami]$_{\text{NP(INST)}}$ [z cyfrowym światem]$_{\text{PREPNP(Z,INST)}}$.

‘Powerful [computers]$_{\text{SUBJ}}$ [link]$_{\text{VERB}}$ [the company]$_{\text{OBJ}}$ [with the digital world]$_{\text{PREPNP(Z,INST)}}$ using [optical fiber]$_{\text{NP(INST)}}$.’

Building Walenty phrase types

Nominal and verbal rules use only **np**, **prepnp**, and **comprepnp** phrases:

- np(**case**)
- prepnp(**prep, case**)
- comprepnp(**complex preposition**)

Where:

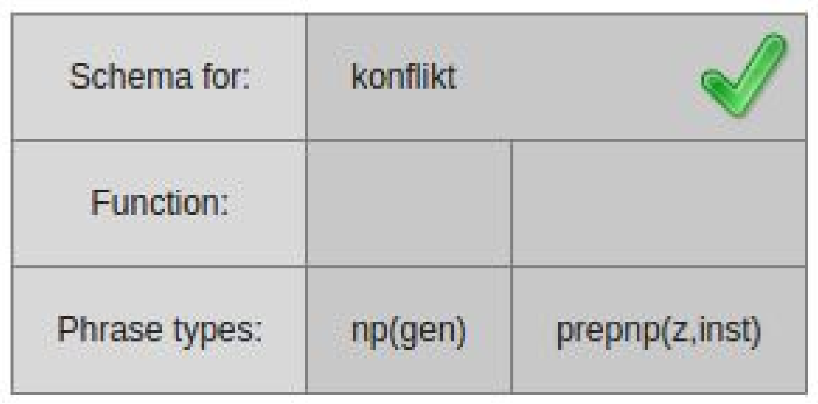
- **case** is case of nominal or prepositional-nominal group head detected by Spejd
- **prep** is preposition word tagged by Spejd as Prep, starting detected prepositional-nominal group
- **complex preposition** is word tagged as Prep but consisting of more than one segment

Nominal realizations (merging)

Od tamtego czasu miał miejsce [konflikt]$_{\text{NOUN}}$ [polskiego ambasadora]$_{\text{NP(GEN)}}$ [z polskim księdzem]$_{\text{PREPNP(Z,INST)}}$.

’Since then there was [a conflict]$_{\text{NOUN}}$ [of the Polish ambassador]$_{\text{NP(GEN)}}$ [with the Polish priest]$_{\text{PREPNP(Z,INST)}}$.’

+



Schema for:	konflikt ✓	
Function:		
Phrase types:	np(gen)	prepnp(z,inst)

=

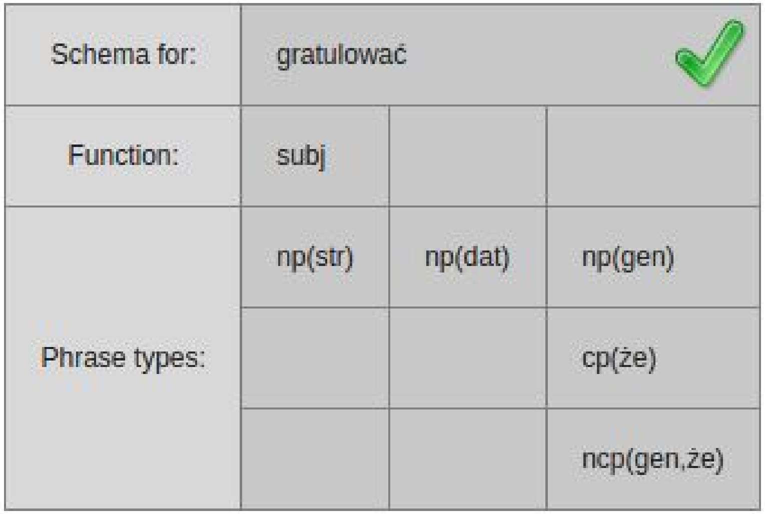
[konflikt polskiego ambasadora z polskim księdzem]
‘[a conflict of the Polish ambassador with the Polish priest]’

Verbal realizations (cleaning)

[Gratuluję]$_{\text{VERB}}$ [Włochom]$_{\text{NP(DAT)}}$ [awansu]$_{\text{NP(GEN)}}$.

’I [congratulate]$_{\text{VERB}}$ [the Italians]$_{\text{NP(DAT)}}$ on their [promotion]$_{\text{NP(GEN)}}$.’

+



Schema for:	gratulować ✓		
Function:	subj		
Phrase types:	np(str)	np(dat)	np(gen)
			cp(że)
			ncp(gen,że)

≠

[Włochom awansu]

‘[the Italians on their promotion]’

Secondary prepositions and phraseological compounds (cleaning)

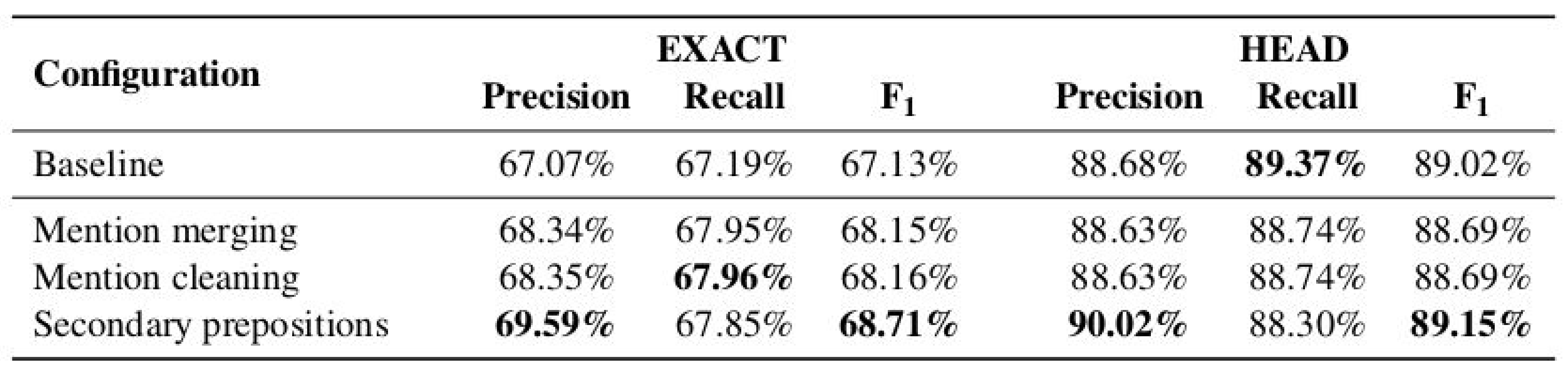
Removing mentions being part of frazeos:

- particle-adverbs (Qub), e.g. *bez wątpienia* ‘without a doubt’
- secondary prepositions (Prep), e.g. *na bazie* ‘based on’
- adverbs (Adv), e.g. *w lot* ’immediately’
- interjections (Interj), e.g. *broń Boże* ’heaven forbid’
- adjectives (Adj), e.g. *na poziomie* ’ambitious’
- conjunctions (Conj), e.g. *przy czym* ’at the same time’
- compounds (Comp), e.g. *w miarę jak (słuchali)* ’as (they listened)’

Polish Coreference Corpus (PCC)

- built upon the National Corpus of Polish
- about 1900 documents from 14 text genres
- about 540K tokens, 180K mentions and 128K coreference clusters
- each text is a 250–350 word sample consisting of full subsequent paragraphs extracted from a larger text
- a smaller subset of long texts (21), 1000 to 4000 segments per text
- nominal, pronominal, and zero mentions

Mention detection evaluation



Configuration		EXACT			HEAD	
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	67.07%	67.19%	67.13%	88.68%	**89.37%**	89.02%
Mention merging	68.34%	67.95%	68.15%	88.63%	88.74%	88.69%
Mention cleaning	68.35%	**67.96%**	68.16%	88.63%	88.74%	88.69%
Secondary prepositions	**69.59%**	67.85%	**68.71%**	**90.02%**	88.30%	**89.15%**

- Precision, recall and F-measure were calculated using Scoreference
- Two alternative mention detection scores: EXACT boundary match and HEAD match.

Future plans

- analyse how other types of phrases intervene in the process of mention construction
- use dependency parser for mention detection instead of Spejd or try to use them both at a time
- check how mention detection score is rising with Walenty expansion (particularly with new noun entries)

Thank you...