Rozpoznawanie wyrażeń temporalnych

oraz opisów sytuacji w dokumentach tekstowych dla języka polskiego

Jan Kocoń

Promotor: prof. dr hab. inż. Halina Kwaśnicka
Promotor pomocniczy: dr inż. Maciej Piasecki
Katedra Inteligencji Obliczeniowej

9 kwietnia 2018

Politechnika
Wrocławska

Plan prezentacji

- Motywacje
- Potok przetwarzania
- Wyrażenia temporalne
- Opisy sytuacji
- Rozpoznawanie wyrażeń językowych
- Autorska metoda SFT (ang. *Selection of Feature Templates*)
- Autorskie rozwiązania pomocnicze:
  - Metoda GDD (ang. *Generating of Domain Dictionaries*)
  - Metoda CVV (ang. *Cross-validation Verification*)
  - Metoda CPR (ang. *Cascade of Partial Rules*)
- Wyniki
- Dalsze kierunki badań

Motywacje

- Pomysł: rezultat prac nad rozpoznawaniem nazw własnych
  - Metoda NER [1, 2] poprawiła skuteczność metody QA dla PL
  - Język polski - wyzwanie dla metod NLP
  - Brak wielu metod podstawowych dla języka polskiego
- Brakujące metody IE dla języka polskiego
  - Rozpoznawanie wyrażeń temporalnych i opisów sytuacji: nadzieja na dalszą poprawę systemów QA
  - Ówczesny trend światowy – SemEval 2013 [3]
  - Intuicje potwierdzone wynikami badań – EMNLP 2017 [4]
- Środki na rozwój metod podstawowych NLP – CLARIN-PL

[1] Marcin Walas i Krzysztof Jassem. "Named entity recognition in a Polish question answering system". W: *Intelligent Information Systems* (2010), s. 181–192

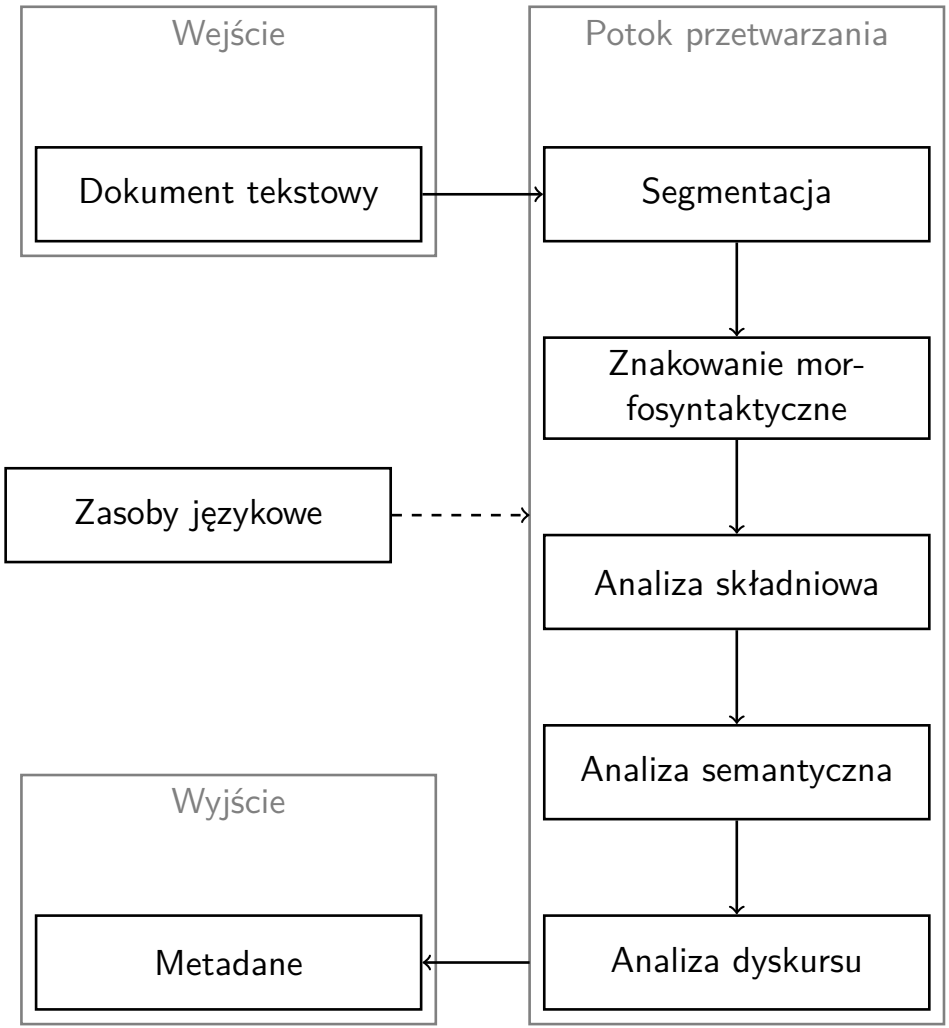
[2] Michał Marcińczuk i in. "Evaluation of baseline information retrieval for Polish open-domain Question Answering system". W: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria, wrz. 2013, s. 428–435

[3] Naushad UzZaman i in. "Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations". W: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. T. 2. 2013, s. 1–9

[4] Yuanliang Meng, Anna Rumshisky i Alexey Romanov. "Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture". W: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, wrz. 2017, s. 887–896

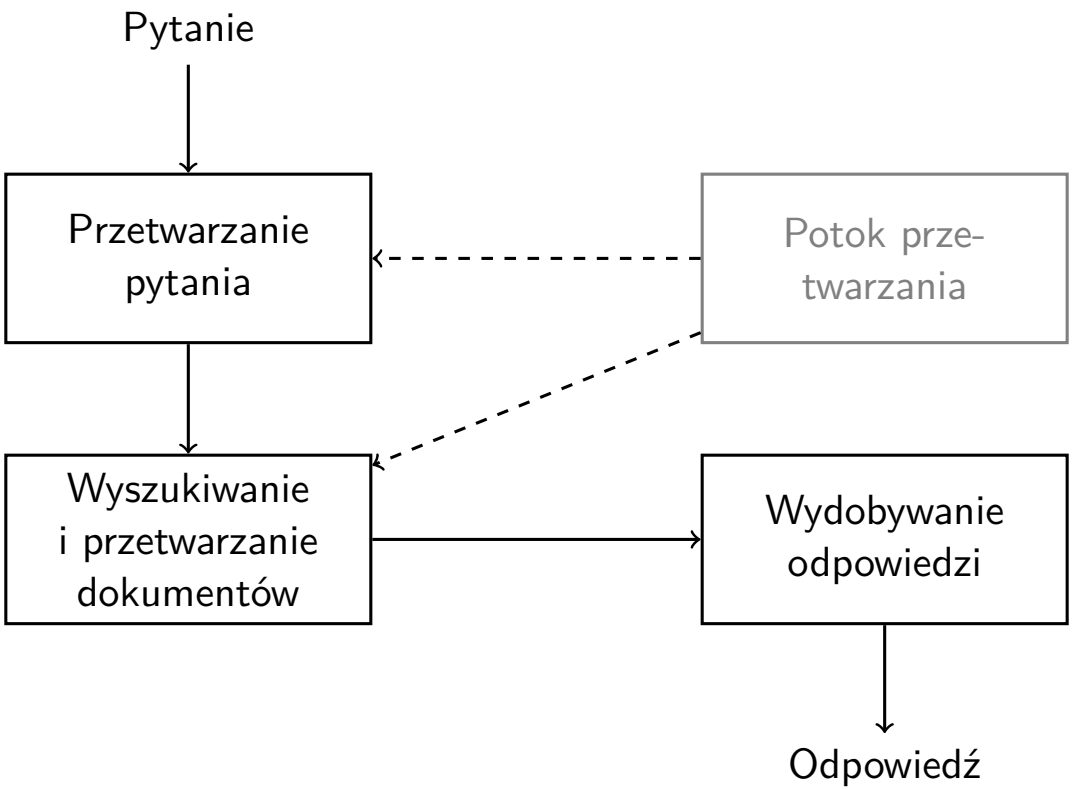
Potok przetwarzania tekstu

Rysunek 1: Typowe etapy przetwarzania tekstu w języku naturalnym w teoretycznym systemie informatycznym, który w pełni analizuje wypowiedź językową w kontekście. Diagram opracowano na podstawie opisu etapów zaczerpniętego z publikacji [5].



[5] Daniel Jurafsky i James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN: 0130950696

Potok przetwarzania tekstu



Rysunek 2: Przykładowa architektura systemu do odpowiadania na pytania, która wykorzystuje przedstawiony potok przetwarzania tekstu. Diagram opracowano na podstawie opisu etapów zaczerpniętego z publikacji [6].

[6] PM Athira, M Sreeja i PC Reghuraj. "Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System". W: *International Journal of Web & Semantic Technology* 4.4 (2013), s. 31

TimeML

- Specyfikacja anotacji tekstu wyrażeniami temporalnymi i opisami sytuacji
- Możliwości:
  - Umiejscowienie sytuacji w czasie
  - Szeregowanie sytuacji w odniesieniu do czasu
  - Wnioskowanie na temat czasu trwania sytuacji
- Zaakceptowany jako standard ISO [9]
- Popularny; adaptacje:
  - Opisów sytuacji dla 6 języków
  - Wyrażeń temporalnych dla 13 języków
- Język polski jako odpowiednio 7. [10] i 14. [11] język z adaptacją TimeML.

---

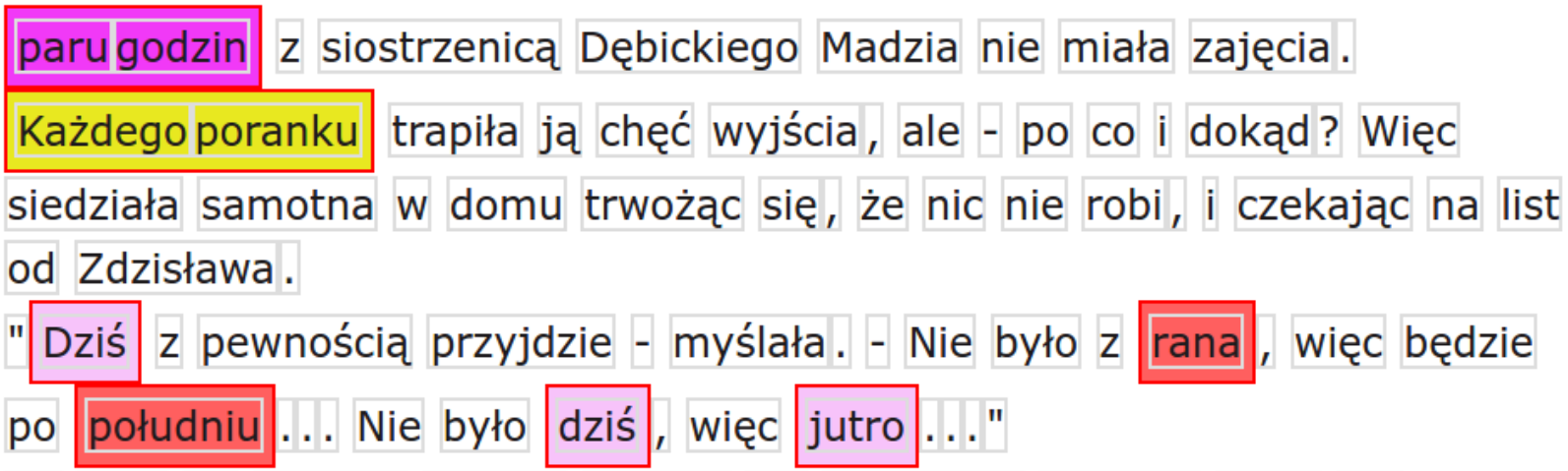
[9] James Pustejovsky i in. "ISO-TimeML: An International Standard for Semantic Annotation". W: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), maj 2010. ISBN: 2-9517408-6-7

[10] Michał Marcińczuk i in. "Towards an Event Annotated Corpus of Polish". W: *Cognitive Studies — Etudes Cognitives* 15 (2015)

[11] Jan Kocoń i in. "Temporal Expressions in Polish Corpus KPWr". W: *Cognitive Studies — Etudes Cognitives* 15 (2015)

Wyrażenia temporalne

- **Kiedy** coś się stało?
- **Jak długo** coś trwało?
- **Jak często** coś się wydarza?
- 4 klasy: data, pora (godzina), trwanie, seria



Rysunek 3: Przykładowe wyrażenia temporalne z Korpusu Języka Polskiego Politechniki Wrocławskiej, KPWr [7] – proza dawna, B. Prus *Emancypantki* (1894)

[7] Bartosz Broda i in. "KPWr: Towards a Free Corpus of Polish". W: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Wyed. Calzolari Nicoletta i in. Istanbul, Turkey: European Language Resources Association (ELRA), 25 maj. 23. ISBN: 978-2-9517408-7-7

Wyrażenia temporalne – liczba anotacji

	Aq	TB	T3P	T3Sil	T3Sp	$\sum$ *inne*	KPWr
Język	EN	EN	EN	EN	SP	EN&SP	PL
Dokumenty	73	183	21	2452	178	2907	1635
Tokeny	33973	61418	6375	666309	67810	835885	447576
Tok./dok.	465	336	303	272	381	288	273
data	502	1177	116	11133	924	13852	4391
godzina	68	44	4	2644	57	2817	928
trwanie	68	189	34	1346	251	1888	653
seria	14	16	4	68	37	139	144
$\sum$ *wszystkie*	652	1426	158	15191	1269	18696	6116
data[%]	76,99	82,54	73,42	73,29	72,81	**74,09**	**71,80**
godzina[%]	10,43	3,09	2,53	17,41	4,49	**15,07**	**15,17**
trwanie[%]	10,43	13,25	21,52	8,86	19,78	**10,10**	**10,68**
seria[%]	2,15	1,12	2,53	0,45	2,92	**0,74**	**2,35**

Tablica 1: Rezultat analizy danych udostępnionych przez organizatorów warsztatów SemEval2013 w porównaniu do danych zawartych w korpusie KPWr.

Wyrażenia temporalne – analiza zgodności

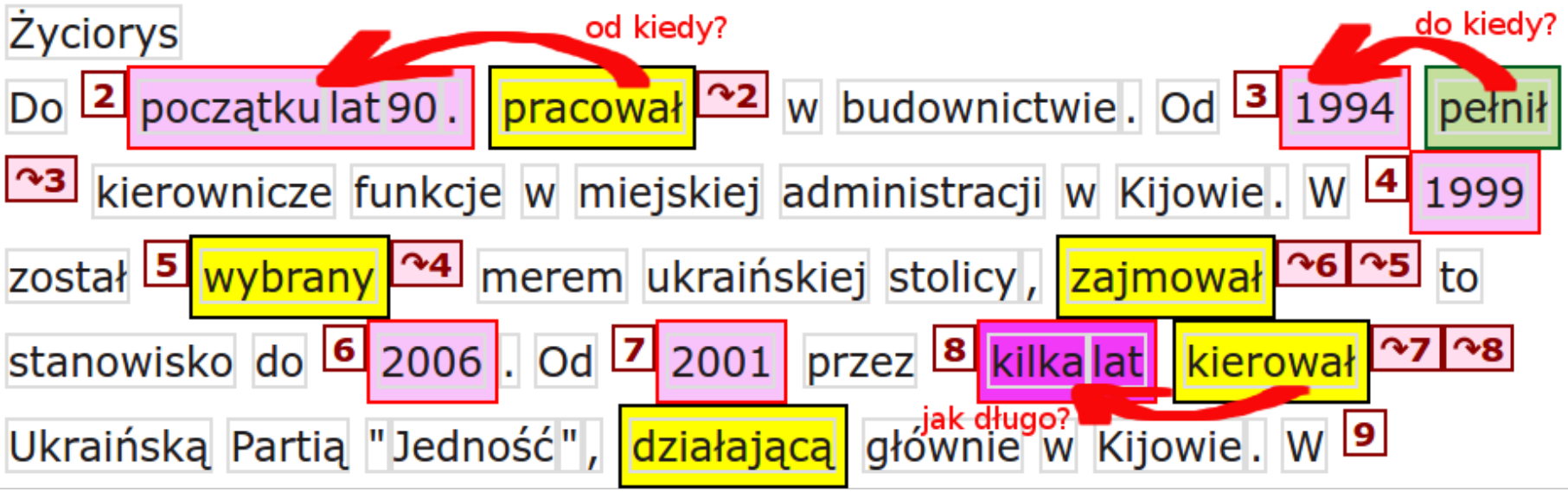
Klasa	1 i 2	tylko 1	tylko 2	PSA [%]
data	182	12	22	91,46
godzina	28	13	8	72,73
trwanie	13	3	4	78,79
seria	6	2	9	52,17
$\sum$	229	30	43	**86,25**

Tablica 2: Wartość współczynnika PSA (ang. *Positive Specific Agreement*) [8] wyznaczonej na podzbiorze losowo wybranych 100 dokumentów z KPWr, oznaczonym niezależnie przez dwóch ekspertów dziedzinowych. *1 i 2* oznacza liczbę anotacji wprowadzonych przez obu ekspertów. *Tylko 1* jest liczbą anotacji wprowadzonych tylko przez pierwszego eksperta, a *tylko 2* jest liczbą anotacji wprowadzonych tylko przez drugiego eksperta.

[8] George Hripcsak i Adam S. Rothschild. "Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval." W: *JAMIA* 12.3 (2005), s. 296–298. DOI: 10.1197/jamia.M1733. URL: http://dblp.uni-trier.de/db/journals/jamia/jamia12.html#HripcsakR05

Opisy sytuacji

- Źródło informacji o zmianach w świecie
- Istotne jest umiejscowienie sytuacji w czasie
- 8 klas opisów sytuacji



Rysunek 4: Przykładowe opisy sytuacji i wyrażenia temporalne połączone relacjami, zaczerpnięte z KPWr – Wikipedia, Ołeksandr Omelczenko (2007)

Opisy sytuacji – przykłady

- Raportowanie: mówić, opowiadać, wyjaśniać, przedstawiać.
- Percepcja: widzieć, ujrzeć, usłyszeć, słuchać.
- Aspektowość: rozpocząć, ponowić, przestać, zakończyć.
- Akcja intensjonalna: spróbować, odwlekać, unikać, prosić.
- Stan intensjonalny: myśleć, wątpić, pragnąć, planować.
- Stan: stać, istnieć.
- Akcja: budować, tańczyć, skakać, biegać.
- Pomocniczy: przeprowadzić, dokonać, spowodować, zadać.

Opisy sytuacji – liczność

Klasa	Liczność	Część całości [%]
akcja	12960	75,34
aspektowość	318	1,85
akcja intensjonalna	720	4,19
stan intensjonalny	1208	7,02
pomocniczy	173	1,01
percepcja	151	0,88
raportowanie	346	2,01
stan	1326	7,71
SUMA	17202	100,00

Tablica 3: Liczność wystąpień opisów sytuacji dla poszczególnych klas w Korpusie Języka Polskiego Politechniki Wrocławskiej (540 dokumentów).

Opisy sytuacji – analiza zgodności

Klasa	A i B	Tylko A	Tylko B	*PSA*[%]
Bez uwzględniania klas	2427	140	225	**93,01**
Z klasami	1856	346	430	**82,71**
akcja	1531	253	198	87,16
stan	135	45	81	68,18
percepcja	23	1	6	86,79
raportowanie	21	18	14	56,76
aspektowość	26	2	8	83,87
akcja intensjonalna	20	8	57	38,10
stan intensjonalny	88	13	48	74,26
pomocniczy	12	6	18	50,00

Tablica 4: Zgodność w oznaczaniu sytuacji pomiędzy dwoma ekspertami (A oraz B) – druga iteracja. Użyto współczynnika PSA.

Rozpoznawanie wyrażeń językowych

- Podejścia: regułowe, oparte o nadzorowane uczenie, hybrydowe
- Wnioski z SemEval 2013 [3]: najlepsze wyniki rozpoznawania miały metody wykorzystujące warunkowe pola losowe (ang. *Conditional Random Fields [12], CRF*).

anotacja	Wojna	wybuchła	1	września	1939	po	ataku	III	Rzeszy
$\vec{y}$	O	O	B-data	I-data	I-data	O	O	O	O
1) TP: $\vec{y_1}$	O	O	B-data	I-data	I-data	O	O	O	O
2) FP&FN: $\vec{y_2}$	O	B-data	I-data	I-data	O	O	O	O	O
3) 2×FP&FN: $\vec{y_3}$	O	O	B-data	O	B-data	O	O	O	O
4) FN: $\vec{y_4}$	O	O	O	O	O	O	O	O	O

Tablica 5: Anotacja wzorcowa i przykładowe wyniki metody.

[3] Naushad UzZaman i in. "Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations". W: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. T. 2. 2013, s. 1–9

[12] John D. Lafferty, Andrew McCallum i Fernando C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". W: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, s. 282–289. ISBN: 1-55860-778-1. URL: http://dl.acm.org/citation.cfm?id=645530.655813

Liniowe warunkowe pola losowe

Niech będą dane:

- $\vec{\lambda}$ – wagi funkcji cech
- x – macierz obserwacji (np. zdanie zawierające n wyrazów opisanych k cechami)
- $\vec{y}$ – wektor stanów (np. klasy)
- $\mathfrak{Y}$ – zbiór możliwych stanów (klas)
- $f_i\left(y_{j-1}, y_j, x, j\right), i \in \{1, \ldots, m\}$ – funkcje cech

Model liniowych warunkowych pól losowych:

$$p_{\vec{\lambda}}(\vec{y}|x) = \frac{1}{Z_{\vec{\lambda}}(x)} \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\left(y_{j-1}, y_j, x, j\right)\right). \qquad (1)$$

Funkcja normalizacji:

$$Z_{\vec{\lambda}}(x) = \sum_{\vec{y}\in\mathfrak{Y}^n} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\left(y_{j-1}, y_j, x, j\right)\right). \qquad (2)$$

Funkcje cech

j	pozycja	1	2	3	4	5	6	7	8	9
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O	O	O	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po	ataku	III	Rzeszy
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po	atak	iii	rzesza
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep	subst	interj	subst
$\vec{x}_4$	liczba	0	0	1	0	1	0	0	0	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx	xxxxx	XXX	Xxxxxx

Tablica 6: Zdanie opisane cechami, zawierające wyrażenie temporalne.

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0					

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0	0				

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0	0	0			

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0	0	0	1		

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0	0	0	1	0	

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcje cech

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_2$	lemat	wojna	wybuchnąć	1	wrzesień	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0
$\vec{x}_5$	struktura	Xxxxx	xxxxxxxx	d	xxxxxxxx	dddd	xx
f_1	wartość	0	0	0	1	0	0

$$f_1(y_{j-1}, y_j, x, j) = \begin{cases} 1: & y_{j-1} = \text{B-data} \ \& \\ & y_j = \text{I-data} \ \& \\ & x_{4,j-2} = 0 \ \& \\ & x_{4,j-1} = 1 \ \& \\ & x_{5,j+1} = \text{dddd} \\ 0: & \text{w przeciwnym przypadku.} \end{cases}$$

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x, j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathcal{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]

Niech $g_1(x, j) = \{x_{3,j}\}$, $g_2(x, j) = \{x_{4,j-1}\}$

Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x,j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathfrak{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]
Niech $g_1(x,j) = \{x_{3,j}\}$, $g_2(x,j) = \{x_{4,j-1}\}$
Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$
$f_1(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{start} \wedge y_j = \text{O} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = \text{null}$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x,j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathcal{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]
Niech $g_1(x,j) = \{x_{3,j}\}$, $g_2(x,j) = \{x_{4,j-1}\}$
Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$
$f_1(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{start} \wedge y_j = \text{O} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = \text{null}$
$f_2(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{O} \wedge x_{3,j} = \text{praet} \wedge x_{4,j-1} = 0$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x,j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathfrak{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]
Niech $g_1(x,j) = \{x_{3,j}\}$, $g_2(x,j) = \{x_{4,j-1}\}$
Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$
$f_1(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{start} \wedge y_j = \text{O} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = \text{null}$
$f_2(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{O} \wedge x_{3,j} = \text{praet} \wedge x_{4,j-1} = 0$
$f_3(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{B-data} \wedge x_{3,j} = \text{adj} \wedge x_{4,j-1} = 0$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x,j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathcal{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]
Niech $g_1(x,j) = \{x_{3,j}\}$, $g_2(x,j) = \{x_{4,j-1}\}$
Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$
$f_1(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{start} \wedge y_j = \text{O} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = \text{null}$
$f_2(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{O} \wedge x_{3,j} = \text{praet} \wedge x_{4,j-1} = 0$
$f_3(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{B-data} \wedge x_{3,j} = \text{adj} \wedge x_{4,j-1} = 0$
$f_4(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{B-data} \wedge y_j = \text{I-data} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = 1$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Szablony funkcji cech (SFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$g(x,j) : \mathcal{F} = \{\forall y, \dot{y} \in \mathcal{Y}. \quad : f_k(y, \dot{y}, g)\}$ [13]
Niech $g_1(x,j) = \{x_{3,j}\}$, $g_2(x,j) = \{x_{4,j-1}\}$
Złożenie SFC: $g_1 \circ g_2 = \{x_{3,j}, x_{4,j-1}\}$
$f_1(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{start} \wedge y_j = \text{O} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = \text{null}$
$f_2(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{O} \wedge x_{3,j} = \text{praet} \wedge x_{4,j-1} = 0$
$f_3(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{O} \wedge y_j = \text{B-data} \wedge x_{3,j} = \text{adj} \wedge x_{4,j-1} = 0$
$f_4(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{B-data} \wedge y_j = \text{I-data} \wedge x_{3,j} = \text{subst} \wedge x_{4,j-1} = 1$
$f_5(y_{j-1}, y_j, x, j) = \{1 : y_{j-1} = \text{I-data} \wedge y_j = \text{I-data} \wedge x_{3,j} = \text{adj} \wedge x_{4,j-1} = 0$

[13] Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007

Zbiory szablonów funkcji cech (ZSFC)

j	pozycja	1	2	3	4	5	6
$\vec{y}$	klasa	O	O	B-data	I-data	I-data	O
$\vec{x}_1$	wyraz	Wojna	wybuchła	1	września	1939	po
$\vec{x}_3$	cz. m.	subst	praet	adj	subst	adj	prep
$\vec{x}_4$	liczba	0	0	1	0	1	0

$$G(-2, 2, x_3) = \{g_1, g_2, g_3, g_4, g_5\}$$

$$g_1(x, j) = \{x_{3,j-2}\}$$
$$g_2(x, j) = \{x_{3,j-1}\}$$
$$g_3(x, j) = \{x_{3,j}\}$$
$$g_4(x, j) = \{x_{3,j+1}\}$$
$$g_5(x, j) = \{x_{3,j+2}\}$$

Złożoność obliczeniowa

- Czas wnioskowania zależny liniowo od liczby funkcji cech.
- Przykład – rozpoznawanie wyrażeń temporalnych (4 klasy):
  - każda klasa – 2 klasy w formacie IOB, np. *data:B-data, I-data*
  - klasy IOB łącznie: 9;
  - dokumenty: 1500;
  - zdania: 25406,
  - ZSFC: 56;
  - SFC: 280
  - **FC: 3981578**
- Jak w skończonym czasie zredukować zbiór FC?

Cel rozprawy

Opracowanie metody do rozpoznawania wyrażeń temporalnych i opisów sytuacji, wykorzystującej metodę **SFT** do otrzymania takiego zestawu ZSFC, który gwarantuje zachowanie niepogorszonej jakości rozpoznawania przy jednoczesnym zwiększeniu wydajności metody w porównaniu do metody wykorzystującej podstawowy zestaw ZSFC.

- Niepogorszona jakość – brak istotnego statystycznie [14] spadku wartości *miary F* dla rozpoznawania wyrażeń językowych
- Zmniejszenie złożoności obliczeń – istotne statystycznie skrócenie czasu rozpoznawania wyrażeń językowych

[14] Thomas G. Dietterich. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". W: *Neural Computation* 10 (1998), s. 1895–1923

Zadania badawcze

Tablica 6: Wykaz zadań badawczych dotyczących rozpoznawania wyrażeń temporalnych (ZT) i opisów sytuacji (ZS).

ZT1	ZS1	*Opracowanie wytycznych opisu* *
ZT2		Opracowanie wytycznych normalizacji WT
ZT3	ZS2	*Anotacja korpusu* *
ZT4	ZS3	*Analiza zgodności* *
ZT5	ZS4	Opracowanie autorskich cech
ZT6	ZT6	Opracowanie autorskiej metody **SFT**
ZT7	ZS5	Analiza istotności statystycznej (jakość)
ZT8		Opracowanie metody do normalizacji WT
ZT9	ZS6	Analiza istotności statystycznej (wydajność)

* zadania badawcze realizowane głównie przez zespół lingwistów z grupy G4.19 pod kierownictwem dra Marcina Oleksego

Metoda selekcji i tworzenia funkcji cech [15]

1. Wejście: Zbiór funkcji cech C, zbiór dokumentów D;
2. $M = \emptyset$;
3. Dopóki $C \neq \emptyset$ i poprawa jakości:
   1. $f^* = \underset{f \in C}{\text{argmax}} \left(EVALCV10(CRF(M \cup \{f\}, D)) \right)$;
   2. $M = M \cup \{f^*\}$ oraz $C = C - \{f^*\}$;
   3. $C = C \cup \{f^* \circ f | \forall f \in M \wedge f, f^* \text{ nie są złożone}\}$;
4. Wyjście: M;

Złożoność obliczeniowa: $O(n^4), n = |C|$.
Przykładowo dla $n = 10^6$ będzie maksymalnie 10^{24} iteracji.

---

[15] Xiaojin Zhu. "Conditional Random Fields". CS769 Advanced Natural Language Processing. 2010. URL: http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf

Autorska metoda SFT - etap I [16]

1. Wejście: Zestaw ZSFC C, zbiór dokumentów D;
2. $M = \emptyset$;
3. Dopóki $C \neq \emptyset$ i poprawa jakości:
   1. $G^* = \underset{G \in C}{\text{argmax}} \left(EVALCV10(CRF(M \cup \{G\}, D)) \right)$;
   2. $M = M \cup \{G^*\}$ oraz $C = C - \{G^*\}$;
4. $\forall G \in M : G = G \cup \{g_2 \circ g_3\}, \{g_2, g_3\} \subset G$
5. Wyjście: M;

Złożoność obliczeniowa $\frac{1}{2}n^2 = O(n^2), n = |C|$.
Przykładowo dla $n = 60$ będzie maksymalnie 1800 iteracji.

---

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Autorska metoda SFT - etap II [16]

1. Wejście: Zbiór SFC C z etapu I, zbiór dokumentów D;
2. $M = \emptyset$;
3. Dopóki $C \neq \emptyset$ i poprawa jakości:
   1. $g^* = \underset{g \in C}{\operatorname{argmax}} \left(EVALCV10(CRF(M \cup \{g\}, D)) \right)$;
   2. $M = M \cup \{g^*\}$ oraz $C = C - \{g^*\}$;
4. Wyjście: M;

Złożoność obliczeniowa $\frac{1}{2}n^2 = O(n^2), n = |C|$.
Przykładowo dla $n = 300$ będzie maksymalnie 45000 iteracji.

---

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Rozpoznawanie wyrażeń temporalnych – ocena jakości

Porównanie wyników dla ZSFCB (bazowy zestaw ZSFC) z zestawem ZSFCBAS2 (zestaw bazowy + autorskie ZSFC (ZSFCBA), następnie zredukowany przy pomocy metody SFT)

Klasa	ZSFCB F[%]	ZSFCBA F[%]	ZSFCBAS2 F[%]
data	89.54	89.81	89.78
godzina	60.43	63.41	65.52
trwanie	66.84	69.73	71.36
seria	48.28	48.28	48.28

Tablica 7: Wyniki oceny jakości rozpoznawania wyrażeń temporalnych na zbiorze testowym [16] – porównanie kolumn ZSFCB i ZSFCBAS2.

Wniosek: Brak istotnych statystycznie różnic w jakości

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Rozpoznawanie wyrażeń temporalnych – ocena wydajności

Porównanie wyników czasu przetwarzania – analogicznie.

ZSFCB t[h]	ZSFCBA t[h]	ZSFCBAS1 t[h]	ZSFCBAS2 t[h]
91,27	**153,18**	**45,52**	**36,84**

Tablica 8: Czas przetwarzania (w godzinach) dla różnych zestawów ZSFC – przetwarzanie 4,2 mld tokenów (KGR10) na serwerze z procesorem Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz (8 rdzeni, 16 wątków HT) i 122GB RAM. **Różnice** pomiędzy **ZSFCB** a **pozostałymi wynikami**, które zostały oznaczone pogrubioną czcionką, są **istotne statystycznie** [16].

Wniosek: Istotny statystycznie spadek czasu przetwarzania

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Rozpoznawanie opisów sytuacji – ocena jakości

Klasa	ZSFCBA F[%]	ZSFCBAS1 F[%]	ZSFCBAS1P F[%]
akcja	81,55	81,57	**83,18**
aspektowość	59,03	55,03	**70,68**
akcja intensjonalna	38,95	38,82	**49,79**
stan intensjonalny	79,81	80,09	**81,20**
pomocniczy	15,38	18,07	**24,28**
percepcja	45,57	37,33	**67,34**
raportowanie	60,40	55,35	59,61
stan	60,03	58,74	**65,00**
$\sum$	77,20	77,07	**79,32**

Tablica 9: Wyniki (miara F) dla każdej **klasy** sytuacji (CV10). Analiza wpływu autorskiej metody GDD[1]: ZSFCBAS1P [17]

Wniosek: SFT+GDD: istotna statystycznie poprawa jakości

[1]Tworzenie słowników dziedzinowych (ang. *Generating of Domain Dictionaries*)

[17] Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016

Rozpoznawanie opisów sytuacji – ocena wydajności cz.1

ZSFC	Czas [s]	Miara F [%]
ZSFCBA	194	77,40
ZSFCBAS1	**58**	77,81
ZSFCBAS1P	**68**	**79,82**

Tablica 10: Porównanie czasów przetwarzania oraz jakości w metodzie rozpoznawania opisów sytuacji na zbiorze testowym. Pogrubioną czcionką oznaczono te wyniki, dla których różnice między modelem wykorzystującym ZSFCBA a pozostałymi są istotne statystycznie [17].

Wniosek: Istotny statystycznie spadek czasu przetwarzania

[17] Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016

Rozpoznawanie opisów sytuacji – ocena wydajności cz.2

ZSFC	Czas [h]
ZSFCBA	128,40
ZSFCBAS1	**39,34**
ZSFCBAS1P	**44,75**

Tablica 11: Czas przetwarzania (w godzinach) dla różnych zestawów ZSFC – przetwarzanie 4,2 mld tokenów (KGR10) na serwerze z procesorem Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz (8 rdzeni, 16 wątków HT) i 122GB RAM. **Różnice** pomiędzy **ZSFCB** a **pozostałymi wynikami**, które zostały oznaczone pogrubioną czcionką, są **istotne statystycznie** [17].

Wniosek: Istotny statystycznie spadek czasu przetwarzania

[17] Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016

Autorskie rozwiązania pomocnicze

- Metoda GDD (ang. *Generating of Domain Dictionaries*) – metoda tworzenia słowników dziedzinowych [17].
- Metoda CVV (ang. *Cross-validation Verification*) – metoda poprawy danych w korpusie w oparciu o wyniki walidacji krzyżowej [18].
- Metoda CPR (ang. *Cascade of Partial Rules*) – dokładniejsza metoda normalizacji wyrażeń temporalnych [18].

---

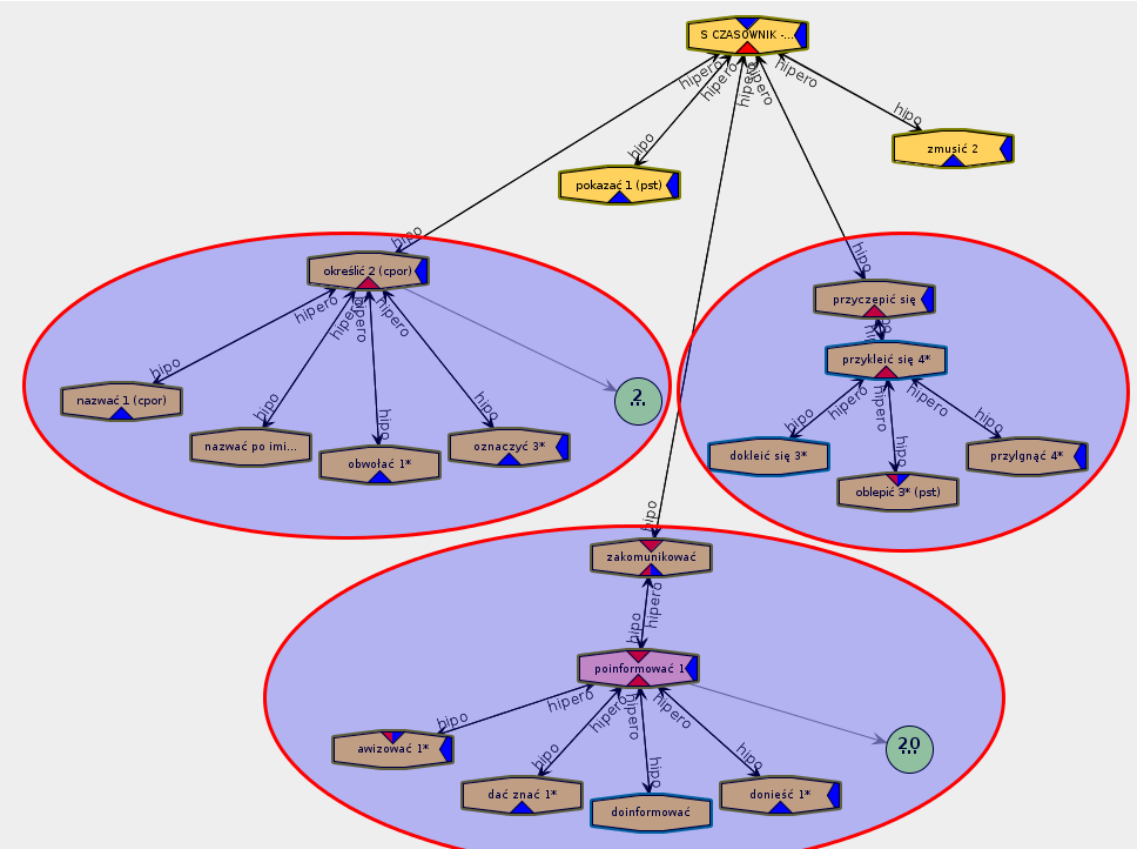
[17] Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016

[18] Jan Kocoń i Michał Marcińczuk. "Improved Recognition and Normalisation of Polish Temporal Expressions". W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 387–393

Metoda GDD

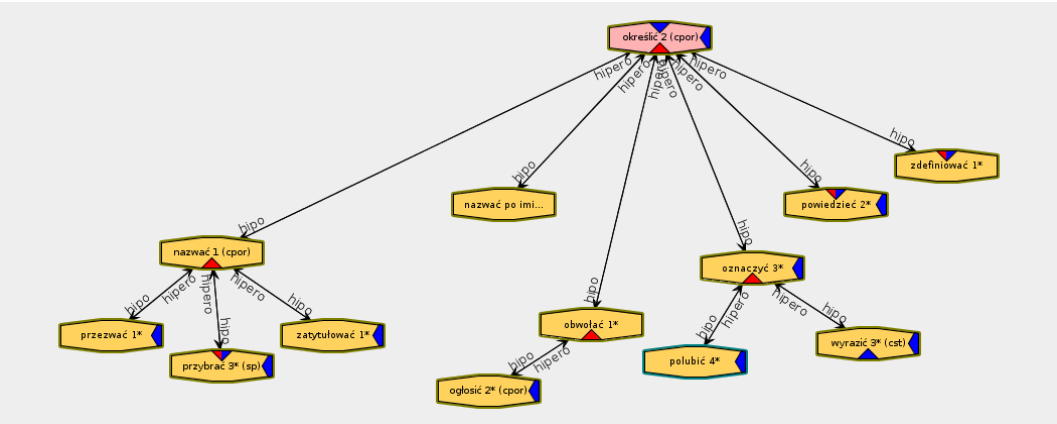
Hipoteza: generalizacja pewnych wyrazów (np. opisów sytuacji) w podzbiorze dokumentów z korpusu (podkorpusie) umożliwia zlokalizowanie tych synsetów w Słowosieci, dla których możliwe jest odtworzenie słowników opisujących obserwowane zjawisko i ułatwia rozróżnienie pomiędzy klasami semantycznymi wyrazów (np. klasami opisów sytuacji) dla innej części tego podzbioru.

Metoda GDD



Rysunek 5: Każdy synset ze Słowosieci jest rozszerzony o wszystkie jednostki leksykalne z wszystkich jego hiponimów – reprezentuje "poddrzewo", którego jest korzeniem. Zbiór jednostek leksykalnych z takiego "poddrzewa" jest kandydatem do włączenia do wyjściowego słownika dziedzinowego.

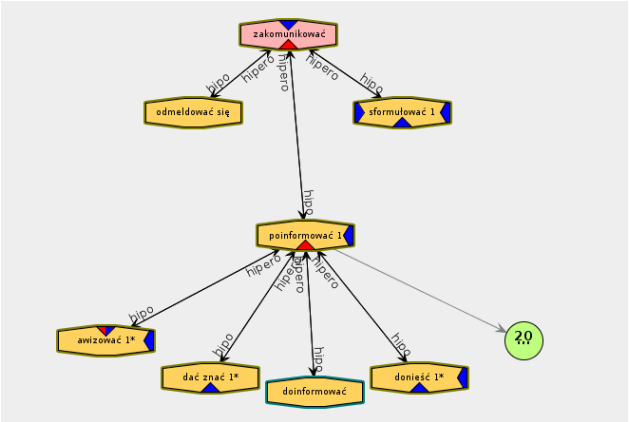
Metoda GDD



Rysunek 6: Przykładowe poddrzewo 1.

- Wektor reprezentujący przynależność wyrazów z podkorpusu do wybranej klasy, np. *raportowanie*:
  $\vec{w} = [0, 1, 0 \ldots 0, 0, 1, 0 \ldots 1, 0, 1, 0 \ldots 0]$
- Lematy poddrzewa 1 z rysunku 6: $v_1 =$ {określić, nazwać, przezwać, przybrać, zatytułować, obwołać, polubić, ...}.
- Wektor poddrzewa 1: $\vec{a}^{v_1} = [0, 1, 0, 0, 0, 1, 0, 0 \ldots]; |\vec{a}^{v_1}| = |\vec{w}|$

Metoda GDD



Rysunek 7: Przykładowe poddrzewo 2., v_2 ={donieść, sformułować, ...}

- Kandydatami są takie poddrzewa, dla których wartość $pearson(\vec{a}^v, \vec{w})$ jest największa.
- Wektor poddrzewa z rys.6: $\vec{a}^{v_1} = [0, 1, 0, 0, 0, 1, 0, 0...]$
- Wektor poddrzewa z rys.7: $\vec{a}^{v_2} = [1, 0, 0, 0, 1, 1, 0, 1...]$
- Połączone ww. wektory:$\vec{a}^{v_1}|\vec{a}^{v_2} = [1, 1, 0, 0, 1, 1, 0, 1, ...]$
- Iteracyjna metoda łączenia najlepszych kandydatów: $\max\left(pearson(\vec{a}^{v_1}, \vec{w}), pearson(\vec{a}^{v_2}, \vec{w})\right) < pearson(\vec{a}^{v_1}|\vec{a}^{v_2}, \vec{w})$

Metoda GDD

Przykłady 3-4 wybranych wyrazów ze słowników dziedzinowych dla każdej klasy (liczność słownika):

- Opisy sytuacji:
  - Raportowanie (102): wygłaszać, podsumować, podpytywać
  - Percepcja (45): usłuchać, oglądanie, widzieć, przyglądać
  - Stan intensjonalny (242): obawiać, planować, móc
  - Akcja intensjonalna (244): kazać, namówić, obwiniać
  - Akcja (25737): denerwować, obtłuc, remont, zaniknięcie
- Listy pożegnalne samobójców [19]:
  - Prawdziwe (5281): tęgość, lizusostwo, uległość, alkohol
  - Sfałszowane (780): dobroduszność, obłudnik, truchleć, mięczak
  - Pozostałe (36111): kserografia, mielonka, umilać, balsamowy

[19] Maciej Piasecki, Ksenia Młynarczyk i Jan Kocoń. "Recognition of genuine Polish suicide notes." W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 583–591

Metoda GDD – podsumowanie

- Istotna poprawa jakości rozpoznawania opisów sytuacji [17].
- Istotna poprawa jakości rozpoznawania listów pożegnalnych samobójców, w tym rozróżniania między prawdziwymi a sfałszowanymi listami [19].

---

[17] Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016

[19] Maciej Piasecki, Ksenia Młynarczyk i Jan Kocoń. "Recognition of genuine Polish suicide notes." W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 583–591

Metoda CVV

- Wejście: korpus KPWr anotowany wyrażeniami temporalnymi, podzielony na 10 części.
- Model: CRF + ZSFCBAS2 (rozpoznawanie w. temp.).
- Wyjście: wyniki 10-krotnej walidacji krzyżowej.
- Dalsze kroki: ręczna weryfikacja przypadków niezgodności rozpoznawania przez lingwistę:
  - Fałszywie dodatnie (FP): 757
  - Fałszywie ujemne (FN): 1126
  - Łącznie do sprawdzenia: 1883
  - Łącznie wszystkich anotacji: 6135

Metoda CVV

Klasa	Przed korektą	Po korekcie	Nowe	Zmienione	SUMA
data	4365	4488	123	29	152
godzina	657	684	27	5	33
trwanie	954	1001	47	10	57
seria	159	185	26	3	29
SUMA	6135	6358	223	47	270

Tablica 12: Liczność anotacji wyrażeniami temporalnymi w KPWr przed i po korekcie. 270 ze 1883 przypadków (ponad 14%) wymagało dokonania korekty w korpusie KPWr.

Normalizacja wyrażeń temporalnych

- Przekształcenie wyrażenia temporalnego do postaci zrozumiałej przez komputer
- Dwa etapy normalizacji:
  - Normalizacja lokalna (ang. *Local Value, LVAL*) – przekształcenie wyrażenia bez brania pod uwagę kontekstu dokumentu (LTIMEX [20])
  - Normalizacja globalna (ang. *Value, VAL*) – uwzględniająca kontekst całego dokumentu (TimeML, TIMEX3 [21])
- Konieczność dostosowania wytycznych normalizacji dla języka polskiego [11] oraz opracowania metody normalizacji [16].

---

[20] Paweł Mazur. "Broad-Coverage Rule-Based Processing of Temporal Expressions". rozprawa doktorska. Politechnika Wrocławska, 2012

[21] Roser Saurí i in. *TimeML Annotation Guidelines, Version 1.2.1*. 2006

[11] Jan Kocoń i in. "Temporal Expressions in Polish Corpus KPWr". W: *Cognitive Studies — Etudes Cognitives* 15 (2015)

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Normalizacja wyrażeń temporalnych

wyrażenie temporalne	LVAL	VAL
dwa dni temu	-0000-00-02	1995-06-04
dwa tygodnie temu	-0000-W02	1995-05-23
sześćdziesiątym ósmym	xx68	1968
8 wieczorem w piątek	xxxx-Wxx-5T20:00	1995-05-26T20:00
o 8 w piątek	xxxx-Wxx-5t08:00	1995-05-26t08:00
następna środa	>D3	1995-05-24
dziewięć miesięcy	P9M	P9M

Tablica 13: Wartości LVAL oraz VAL dla przykładowych wyrażeń temporalnych.

Normalizacja wyrażeń temporalnych

- Podejścia literaturowe: wyłącznie regułowe.
- Stosowane dokładne i specyficzne reguły, opisujące całe wyrażenie temporalne.
- Najlepszy system na SemEval2013: HeidelTime – 326 reguł, VAL F1: 77,61% [3].
- Adaptacja dla języka polskiego w systemie Liner2: 224 reguły, VAL F1: 66,71% [16].
- Dalsze zwiększanie zbioru reguł – bardzo czasochłonne.

---

[3] Naushad UzZaman i in. "Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations". W: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. T. 2. 2013, s. 1–9

[16] Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255

Kaskada reguł częściowych

- Reguły dla klasy *godzina* są często rozszerzeniem reguł dla klasy *data*, można je połączyć.
- Mając rozpoznane wyrażenie temporalne (granice oraz klasę), można niezależnie rozpatrywać składowe wyrażeń.
- Reguły w postaci wyrażeń regularnych, opisujących fragmenty wyrażeń temporalnych
- Reguły przetwarzane w kolejności od najbardziej ogólnych do najbardziej szczegółowych.
- Wiele reguł może być aplikowanych dla jednego wyrażenia temporalnego.
- Reguły mogą nadpisywać składowe rozpoznane przez poprzednie reguły.
- Dla języka polskiego utworzono 167 reguł w systemie Liner2, VAL F1: 77,23% [18].

---

[18] Jan Kocoń i Michał Marcińczuk. "Improved Recognition and Normalisation of Polish Temporal Expressions". W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 387–393

Metoda CVV + CPR – wyniki

Metoda	Rozp.	J.	Luźna F1[%]	Luźna P[%]	Luźna R[%]	Ścisła F1[%]	VAL F1[%]	LVAL F1[%]
HeidelTime	R	A	90,30	93,08	87,68	81,34	77,61	
NavyTime	R	A	90,32	89,36	91,30	79,57	70,97	
ManTime	M	A	89,66	95,12	84,78	74,33	68,97	
SUTime	R	A	90,32	89,36	91,30	79,57	67,38	
ATT	M	A	85,25	98,11	75,36	78,69	65,57	
TIPSem	M	A	84,90	97,20	75,36	81,63	65,31	
ClearTK	M	A	90,23	93,75	86,96	82,71	64,66	
JU-CSE	M	A	86,38	93,28	80,43	75,49	63,81	
KUL	H	A	83,67	92,92	76,09	69,32	62,95	
FSS-TimEx	R	A	85,06	90,24	80,43	49,04	58,24	
Liner2-HeidelTime	M	P	88,50	90,25	86,81	85,06	66,71	75,14
Liner2-CPR	M	P	92,83	94,56	91,15	88,76	77,23	89,23

Tablica 14: Ocena potoku przetwarzania dla metod do rozpoznawania i normalizacji wyrażeń temporalnych (w tym CPR) [18].

Dalsze kierunki badań

Category	branched bi gru-lstm W=1			Liner2 without dictionaries			Liner2 with dictionaries (*)		
	P	R	F1	P	R	F1	P	R	F1
action	84.90	88.33	86.57	82.51	84.90	83.69	82.49	83.87	83.18
aspectual	85.87	72.96	78.67	87.56	60.13	71.29	87.58	59.24	70.68
i_action	66.89	58.58	62.12	67.48	42.54	52.18	63.56	40.92	49.79
i_state	84.35	82.60	83.38	84.35	78.26	81.19	85.19	77.56	81.20
light predicate	62.56	18.63	27.66	45.33	19.88	27.64	56.76	15.44	24.28
perception	85.17	75.61	79.33	97.53	53.02	68.70	85.90	55.37	67.34
reporting	69.29	66.65	67.11	75.00	57.18	64.89	71.13	51.30	59.61
state	73.03	69.09	70.86	71.84	61.15	66.07	68.10	62.17	65.00

Table 5: Comparison of the best performing network architecture against the previously proposed CRF-based approach. Ten-fold cross-validation on the KPWr-540 corpus. (*) Results taken from (Kocoń and Marcińczuk, 2016).

Rysunek 8: Wyniki rozpoznawania opisów sytuacji otrzymane z wykorzystaniem głębokich sieci neuronowych w IPI PAN za pomocą narzędzia DeepEvents[2] [22].

[2] `http://clip.ipipan.waw.pl/DeepEvents`

[22] Łukasz Kobyliński i Michał Wasiluk. "Deep Learning in Event Detection in Polish". W: *w trakcie recenzji*. 2017

Dalsze kierunki badań

- Wykorzystanie głębokich sieci neuronowych do rozpoznawania wyrażeń temporalnych.
- Rozpoznawanie opisów sytuacji z wykorzystaniem głębokich sieci neuronowych: dołączenie do zestawu cech słowników dziedzinowych wygenerowanych na zbiorze uczącym.
- Powtórzenie rozpoznawania opisów sytuacji na pełnym korpusie z wykorzystaniem 10-krotnej walidacji krzyżowej oraz ze słownikami generowanymi na pełnym zbiorze uczącym dla każdej części.

Dziękuję za uwagę

Bibliografia I


Marcin Walas i Krzysztof Jassem. "Named entity recognition in a Polish question answering system". W: *Intelligent Information Systems* (2010), s. 181–192.


Michał Marcińczuk i in. "Evaluation of baseline information retrieval for Polish open-domain Question Answering system". W: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria, wrz. 2013, s. 428–435.


Naushad UzZaman i in. "Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations". W: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. T. 2. 2013, s. 1–9.

Bibliografia II

Yuanliang Meng, Anna Rumshisky i Alexey Romanov. "Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture". W: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, wrz. 2017, s. 887–896.

Daniel Jurafsky i James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN: 0130950696.

Bibliografia III


PM Athira, M Sreeja i PC Reghuraj. "Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System". W: *International Journal of Web & Semantic Technology* 4.4 (2013), s. 31.


Bartosz Broda i in. "KPWr: Towards a Free Corpus of Polish". W: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Wyed. Calzolari Nicoletta i in. Istanbul, Turkey: European Language Resources Association (ELRA), 25 maj. 23. ISBN: 978-2-9517408-7-7.


George Hripcsak i Adam S. Rothschild. "Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval." W: *JAMIA* 12.3 (2005), s. 296–298. DOI: 10.1197/jamia.M1733. URL: http://dblp.uni-trier.de/db/journals/jamia/jamia12.html#HripcsakR05.

Bibliografia IV


James Pustejovsky i in. “ISO-TimeML: An International Standard for Semantic Annotation”. W: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), maj 2010. ISBN: 2-9517408-6-7.


Michał Marcińczuk i in. “Towards an Event Annotated Corpus of Polish”. W: *Cognitive Studies — Etudes Cognitives* 15 (2015).


Jan Kocoń i in. “Temporal Expressions in Polish Corpus KPWr”. W: *Cognitive Studies — Etudes Cognitives* 15 (2015).

Bibliografia V

John D. Lafferty, Andrew McCallum i Fernando C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". W: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, s. 282–289. ISBN: 1-55860-778-1. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

Roman Klinger i Katrin Tomanek. *Classical Probabilistic Models and Conditional Random Fields*. 2007.

Thomas G. Dietterich. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". W: *Neural Computation* 10 (1998), s. 1895–1923.

Bibliografia VI

Xiaojin Zhu. "Conditional Random Fields". CS769 Advanced Natural Language Processing. 2010. URL: http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf.

Jan Kocoń i Michał Marcińczuk. "Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes". W: *Natural Language Engineering* 23.3 (2017), 385–418. DOI: 10.1017/S1351324916000255.

Jan Kocoń i Michał Marcińczuk. "Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents". W: *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*. T. 9924. Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer, wrz. 2016.

Bibliografia VII


Jan Kocoń i Michał Marcińczuk. "Improved Recognition and Normalisation of Polish Temporal Expressions". W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 387–393.


Maciej Piasecki, Ksenia Młynarczyk i Jan Kocoń. "Recognition of genuine Polish suicide notes." W: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 2017, s. 583–591.


Paweł Mazur. "Broad-Coverage Rule-Based Processing of Temporal Expressions". rozprawa doktorska. Politechnika Wrocławska, 2012.


Roser Saurí i in. *TimeML Annotation Guidelines, Version 1.2.1*. 2006.

Bibliografia VIII

Łukasz Kobyliński i Michał Wasiluk. “Deep Learning in Event Detection in Polish”. W: *w trakcie recenzji*. 2017.