Nowe metody ekstrakcji walencji czasowników z tekstów w języku polskim

Łukasz Dębowski, Marcin Woliński
{ldebowsk,wolinski}@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Problem

- Aby przeprowadzić automatyczną analizę syntaktyczną tekstu, potrzeba różnorakich zasobów (gramatyki, słowników).

> Jasio podziękował Marysi za współpracę.
> Marysia odpowiedziała mu, że nie ma za co.

- Jednym z tych zasobów jest słownik walencyjny:

> ...
> *odpowiedzieć*: np(nom), np(dat), ZE
> ...
> *podziękować*: np(nom), np(dat), za+np(acc)
> ...

- Istnieją takie słowniki na papierze – o niejasnej dokładności.
- Czy lepszy słownik można pozyskać z korpusu tekstów?

Ramy walencyjne

Odpowiedź dziwiła Jasia dwa tygodnie.
Jasia dziwiło, że Marysia tak odpowiedziała.
Jasio dziwił się i dziwił.
Jasio dziwił się i Marysi, i odpowiedzi.
Jasio dziwił się także, że się tak długo dziwi.

$$\mathbf{F}(\textit{dziwić}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{ZE}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{sie}\}, \\ \{\text{np(nom)}, \text{sie}, \text{np(dat)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{ZE}\} \end{array} \right\}$$

Walencje są trudne do opisania

1. Nie wszystkie argumenty mogą współwystępować.
2. Trudno wyliczyć wszystkie ramy.
3. Część argumentów można opuścić.
4. Część argumentów bywa wymagana (czasem warunkowo).
5. Rozróżnienie pomiędzy dopełnieniem (argumentem specyficznym dla danego czasownika) a okolicznikiem (argumentem niezależnym od czasownika) jest nieostre:

> Podziękowanie brzmiało dziwacznie.

6. Klasy równoważnych argumentów są zależne od czasownika:

> Jasio dziwił się { że Marysia tak odpowiedziała. / jej odpowiedzi.

Podejście Brenta (1993)

Dane: nieanotowany korpus tekstów,
płytki parser zwracający jednoznaczne analizy zdań.

Wyliczone: $\mathbf{c(v,f)}$ — zliczenie ramy $\mathbf{f}$ z czasownikiem $\mathbf{v}$,
$\mathbf{c(v)} = \sum_{\mathbf{f}} \mathbf{c(v,f)}$ — zliczenie czasownika $\mathbf{v}$.

Uznajemy, że $\mathbf{f}$ jest ramą czasownika $\mathbf{v}$, gdy

$$\sum_{\mathbf{n=c(v,f)}}^{\mathbf{c(v)}} \binom{\mathbf{c(v)}}{\mathbf{n}} \mathbf{p(f)^n (1-p(f))^{c(v)-n}} \leq \mathbf{0.05},$$

gdzie $\mathbf{p(f)}$ jest dobierane:

1. pod nadzorem tak, by zminimalizować liczbę błędów,
2. bez nadzoru tak, by rozkład $\mathbf{B(\cdot, p(f))}$ dopasował się do pierwszego skupienia na histogramie czasowników pogrupowanych wg proporcji $\mathbf{c(v,f)/c(f)}$.

Nasze podejście

1. Użyć parsera Świgra do analizy czystego tekstu.
2. Zredukować lasy analiz do lasów ram walencyjnych.
3. Ujednoznacznić las ram za pomocą nowego algorytmu wyboru EM.
4. Zliczyć wystąpienia ram i czasowników.
5. Zastosować uczenie pod nadzorem do ustalenia zbiorów możliwych i obligatoryjnych argumentów.
6. Użyć formalizmu macierzy współwystąpień i uczenia pod nadzorem do naprawienia zbiorów całych ram.

Parsowanie

Teksty do ekstrakcji walencji pochodziły z Korpusu IPI PAN.

Analizowaliśmy je zmodyfikowaną Świgrą:

- dowolny czasownik mógł mieć $\leq \mathbf{1}$ podmiot i dowolnie wiele innych argumentów.

Do ekstrakcji wybraliśmy zdania:

- długości $\leq \mathbf{15}$ słów,
- analizowane w $\leq \mathbf{1}$ minutę,
- mające $\leq \mathbf{40}$ parsów na zdanie elementarne.

Braliśmy $\leq \mathbf{5000}$ zdań dla jednego czasownika.

Redukcja analiz do ram walencyjnych

Usunięcie niektórych analiz:

- zawierających *to*, *co*, *nic*,
- zawierających skrajnie nieprawdopodobne interpretacje słów.

Przekształcenie pozostałych analiz:

- usunięcie fraz nie będących zależnikami orzeczenia,
- usunięcie zaimka *sam*,
- dodanie podmiotu domyślnego i wyrażonego przez *się*,
- naprawienie dopełniacza negacji,
- oznaczenie niektórych fraz jako okoliczników,
- wykreślenie lematów.

Bank lasów ram walencynych

- Bank zawiera **326 592** zdań elementarnych.
- Świgra rozpatrzyła około **2.6** raza tyle zdań.
- Około **2** miliony słów bieżących.

```
’Kto zastąpi piekarza?’
zastąpić :np:acc: :np:nom:
zastąpić :np:gen: :np:nom:
’Nie płakał na podium.’
płakać :np:nom: :prepnp:na:acc:
płakać :np:nom: :prepnp:na:loc:
```

Algorytm wyboru EM

$\mathbf{Y_i}$ — las ram walencyjnych dla $\mathbf{i}$-tego zdania, $\mathbf{i = 1, 2, ..., M}$.
$\mathbf{p_j^{(n)}}$ — wypadkowa częstość ramy $\mathbf{j}$ w $\mathbf{n}$-tej iteracji, $\mathbf{p_j^{(1)} = 1}$.

$$\mathbf{p_{ji}^{(n)}} = \begin{cases} \mathbf{p_j^{(n)}} / \sum_{\mathbf{j' \in Y_i}} \mathbf{p_{j'}^{(n)}}, & \mathbf{j \in Y_i}, \\ \mathbf{0}, & \text{inaczej}, \end{cases}$$

$$\mathbf{p_j^{(n+1)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{p_{ji}^{(n)}}.$$

wybieranie losowe	akuratność
najkrótszej ramy o największym $\mathbf{p_{ji}^{(n)}}$	**71.3%**
ramy o największym $\mathbf{p_{ji}^{(n)}}$	**70.6%**
najkrótszej ramy	**57.4%**
na ślepo	**46.7%**

Test na 190 zdaniach, 500 prób Monte Carlo, $\mathbf{n = 10}$.

Matematycy są jak Francuzi: cokolwiek im się powie, od razu przekładają to na swój własny język i wówczas staje się to zupełnie czymś innym.

— J. W. Goethe

Krótkie wprowadzenie do ogólnego algorytmu EM

Dempster, Laird, and Rubin (1977):

- $\mathbf{Y}$ — zmienna obserwowana,
- θ — nieznany parametr do oszacowania,
- $\mathbf{P(Y|\theta)}$ — funkcja wiarogodności (rozkład $\mathbf{Y}$ dla każdego θ).

Estymator największej wiarogodności: $\theta_{\mathrm{MLE}} = \mathbf{arg\,max}_{\theta}\, \mathbf{P(Y|\theta)}$.

Gdy nie możemy go policzyć, możemy postąpić tak:

- $\mathbf{Z}$ — dyskretna zmienna ukryta o prostym rozkładzie,
- cross-entropia

$$\mathbf{Q}(\theta', \theta'') = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Z} = \mathbf{z}|\mathbf{Y}, \theta') \log \mathbf{P}(\mathbf{Z} = \mathbf{z}, \mathbf{Y}|\theta''),$$

- obrawszy θ_1, iterujemy $\theta_{n+1} = \mathbf{arg\,max}_{\theta}\, \mathbf{Q}\,(\theta_n, \theta)$.

Łatwo udowodnić, że $\mathbf{P}\,(\mathbf{Y}|\theta_{n+1}) \geq \mathbf{P}\,(\mathbf{Y}|\theta_n)$.

Algorytm wyboru EM w ujęciu probabilistycznym

- $\mathbf{Z = (Z_1, Z_2, ..., Z_M)}$, gdzie $\mathbf{Z_i : \Omega \to J}$ — poprawne ramy.
- $\mathbf{Y = (Y_1, Y_2, ..., Y_M)}$, gdzie $\mathbf{Y_i : \Omega \to 2^J \setminus \emptyset}$ — lasy ram.
- $\theta = \mathbf{(p_j)_{j \in J}}$, gdzie $\mathbf{p_j}$ — p-stwo ramy $\mathbf{j}$,

$$\mathbf{P(Z_i = j|\theta) = p_j.}$$

Algorytm wyboru EM staje się implementacją algorytmu EM, gdy

$$\mathbf{P(Z = (j_1, j_2, ..., j_M), Y|\theta) = \prod_{i=1}^{M} P(Z_i = j_i, Y_i|\theta),}$$

$$\mathbf{P(Y_i = A|Z_i, \theta) = \begin{cases} g(A), & Z_i \in A, \\ 0, & \text{inaczej.} \end{cases}}$$

Na przykład, możemy wziąć $\mathbf{g(A) = q^{|A|-1}(1-q)^{|J|-|A|}}$.

Co maksymalizuje algorytm wyboru EM?

Niezależnie od postaci funkcji $\mathbf{g(\cdot)}$, założenie

$$\mathbf{P(Y_i = A|Z_i, \theta)} = \begin{cases} \mathbf{g(A)}, & \mathbf{Z_i \in A}, \\ \mathbf{0}, & \text{inaczej}, \end{cases}$$

implikuje $\mathbf{P(Y_i|\theta) = g(Y_i)P(Z_i \in Y_i|\theta)}$.

$$\mathbf{P(Y|\theta)} = \prod_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{P(Y_i|\theta)} = \prod_{\mathbf{i=1}}^{\mathbf{M}} \frac{\mathbf{P(Z_i \in Y_i|\theta)}}{\mathbf{g(Y_i)}}.$$

Zatem $\mathbf{L^{(n+1)} \geq L^{(n)}}$ dla $\mathbf{L^{(n)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{\log} \left[\sum_{\mathbf{j \in Y_i}} \mathbf{p_j^{(n)}} \right]$.

Algorytmu wyboru EM można użyć w b. wielu zadaniach NLP.
Np. do dezambiguacji morfologicznej bez próby uczącej:

$\mathbf{Y_i}$ — możliwe interpretacje dla **i**-tego okazu, $\mathbf{i = 1, 2, ..., M}$.
$\mathbf{p_j^{(n)}}$ — częstość interpretacji **j** w **n**-tej iteracji, $\mathbf{p_j^{(1)} = 1}$.

$$\mathbf{p_{ji}^{(n)}} = \begin{cases} \mathbf{p_j^{(n)}} / \sum_{\mathbf{j' \in Y_i}} \mathbf{p_{j'}^{(n)}}, & \mathbf{j \in Y_i}, \\ \mathbf{0}, & \text{inaczej}, \end{cases}$$

$$\mathbf{p_j^{(n+1)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{p_{ji}^{(n)}}.$$

Proto-słownik, próba ucząca i testowa

Proto-słownik (zliczenia ram po ujednoznacznieniu przez EM):

```
'przyłapać' => {
   'np(acc),np(gen),np(nom)' => 1,
   'na+np(loc),np(nom),sie' => 1,
   'na+np(loc),np(gen),np(nom)' => 1,
   'np(acc),np(nom)' => 4,
   'adv,np(nom)' => 1,
   'na+np(loc),np(acc),np(nom)' => 3
}
```

Próba testowa:

- walencje **203** czasowników wg słowników Polańskiego (1980), Świdzińskiego (1994) i Bańki (2000).

Próba ucząca:

- walencje **1166** czasowników wg słownika Świdzińskiego.

Nowy opis walencji

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

Opisujemy **F(v)** w sposób przybliżony za pomocą trzech obiektów:

1. **L(v)** — zbiór argumentów czasownika **v**,
2. **E(v)** — zbiór obligatoryjnych argumentów czasownika **v**,
3. $\mathbf{M(v)} : \mathbf{L(v)} \times \mathbf{L(v)} \rightarrow \{\leftarrow, \rightarrow, \leftrightarrow, \times, \perp\}$
   — macierz współwystąpień argumentów czasownika **v**.

Wartości macierzy współwystąpień słabo zależą od czasownika.

Zbiory argumentów

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

$$\mathbf{L(v)} := \bigcup_{\mathbf{f} \in \mathbf{F(v)}} \mathbf{f}$$

$$\mathbf{E(v)} := \bigcap_{\mathbf{f} \in \mathbf{F(v)}} \mathbf{f}$$

Macierz współwystąpień

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

Niech $\langle \mathbf{a} \rangle := \{\mathbf{f} \in \mathbf{F(v)} \mid \mathbf{a} \in \mathbf{f}\}$ oraz

$$\begin{aligned}
\mathbf{a} \times \mathbf{b} &\iff \langle \mathbf{a} \rangle \cap \langle \mathbf{b} \rangle = \emptyset, && \text{(wykluczanie)} \\
\mathbf{a} \leftrightarrow \mathbf{b} &\iff \langle \mathbf{a} \rangle = \langle \mathbf{b} \rangle, && \text{(współwystępowanie)} \\
\mathbf{a} \rightarrow \mathbf{b} &\iff [\langle \mathbf{a} \rangle \subset \langle \mathbf{b} \rangle \wedge \langle \mathbf{a} \rangle \neq \langle \mathbf{b} \rangle], && \text{(implikacja prawa)} \\
\mathbf{a} \leftarrow \mathbf{b} &\iff [\langle \mathbf{a} \rangle \supset \langle \mathbf{b} \rangle \wedge \langle \mathbf{a} \rangle \neq \langle \mathbf{b} \rangle], && \text{(implikacja lewa)} \\
\mathbf{a} \perp \mathbf{b} &\iff \langle \mathbf{a} \rangle \setminus \langle \mathbf{b} \rangle, \langle \mathbf{a} \rangle \cap \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle \setminus \langle \mathbf{a} \rangle \neq \emptyset. && \text{(niezależność)}
\end{aligned}$$

$$\mathbf{M(v)_{ab}} := \mathrm{R} \iff \mathbf{a}\,\mathrm{R}\,\mathbf{b}$$

Przykład

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

M(*przyłapać*)	np(nom)	np(acc)	sie	na+np(loc)
np(nom)	$\nleftarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
np(acc)	$\uparrow$	$\nleftarrow$	$\times$	$\bot$
sie	$\uparrow$	$\times$	$\nleftarrow$	$\bot$
na+np(loc)	$\uparrow$	$\bot$	$\bot$	$\nleftarrow$

Porównanie macierzy (Bańko vs. Świdziński)

Liczby trójek $\mathbf{(v, a, b)}$ o zadanych wartościach $\mathbf{M(v)_{ab}}$:

		Bań.					
		×	←	→	↔	⊥	N/A
Świ.	×	364	2	2	–	30	724
	←	4	253	1	2	18	176
	→	4	1	253	2	18	176
	↔	–	25	25	410	2	124
	⊥	16	6	6	–	28	38
	N/A	1242	230	230	167	102	

Wysoka zgodność słowników

Liczby trójek $\mathbf{(v, a, b)}$ pojawiających się w parach słowników:

	$\mathbf{\Sigma}$	równe $\mathbf{M(v)_{ab}}$	różne $\mathbf{M(v)_{ab}}$	zgodność
w Bań. i Pol.	1383	1187	196	86%
w Bań. i Świ.	1472	1308	164	89%
w Pol. i Świ.	1449	1283	166	89%

Słowniki walencyjne podają dość odmienne zbiory argumentów dla tych samych czasowników. Jednak łączliwość argumentów wg tych samych słowników jest bardzo podobna.

Najczęstsze wartości macierzy

	np(nom)	np(acc)	advp	np(dat)	np(inst)	w+np(loc)	do+np(gen)	na+np(acc)	z+np(gen)
np(acc)	↑:66%								
adv	↑:94%	↑:57%							
np(dat)	↑:91%	↑:50%	×:72%						
np(inst)	↑:98%	↑:57%	×:85%	×:64%					
w+np(loc)	↑:91%	↑:61%	×:92%	×:71%	×:85%				
do+np(gen)	↑:94%	↑:48%	×:92%	×:80%	×:89%	×:100%			
na+np(acc)	↑:96%	↑:64%	×:96%	×:85%	×:86%	×:90%	×:100%		
z+np(gen)	↑:100%	↑:65%	×:94%	×:91%	×:100%	×:100%	×:92%	×:100%	
ZE	↑:91%	×:76%	×:86%	⊥:64%	×:100%	×:100%	×:80%	×:71%	⊥:100%

W dodatku łączliwość argumentów słabo zależy od czasownika.

Czyszczenie proto-słownika

1. Obliczyć $\mathbf{L(v)}$ i $\mathbf{E(v)}$ z $\mathbf{F(v)}$ w proto-słowniku.
2. Oczyścić $\mathbf{L(v)}$ i $\mathbf{E(v)}$. Następnie zrekonstruować

$$\mathbf{F(v)} := \{(\mathbf{f} \cup \mathbf{E(v)}) \cap \mathbf{L(v)} \mid \mathbf{f} \in \mathbf{F(v)}\}\,.$$

3. Obliczyć $\mathbf{M(v)}$ z bieżącego $\mathbf{F(v)}$.
4. Oczyścić $\mathbf{M(v)}$. Następnie zrekonstruować

$$\mathbf{F(v)} := \left\{ \mathbf{f} \in 2^{\mathbf{L(v)}} \,\middle|\, \begin{array}{l} \forall_{\mathbf{a} \in \mathbf{E(v)}}\, \mathbf{a} \in \mathbf{f}, \\ \forall_{\mathbf{a},\mathbf{b} \in \mathbf{L(v)}}\, \phi(\mathbf{f}, \mathbf{M(v)}, \mathbf{a}, \mathbf{b}) \end{array} \right\},$$

gdzie

$$\phi(\mathbf{f}, \mu, \mathbf{a}, \mathbf{b}) := \begin{cases} \neg(\mathbf{a} \in \mathbf{f} \wedge \mathbf{b} \in \mathbf{f}), & \mu_{\mathbf{ab}} = \times, \\ \mathbf{a} \in \mathbf{f} \iff \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \leftrightarrow, \\ \mathbf{a} \in \mathbf{f} \implies \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \rightarrow, \\ \mathbf{a} \in \mathbf{f} \impliedby \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \leftarrow, \\ \text{prawda}, & \mu_{\mathbf{ab}} = \perp\,. \end{cases}$$

Czyszczenie zbiorów argumentów

Argument **a** jest uznawany za możliwy dla czasownika **v**, jeżeli

$$\mathbf{c(v, a)} \geq \mathbf{p_a c(v)} + \mathbf{1},$$

zaś za obligatoryjny dla czasownika **v**, jeżeli jest możliwy oraz

$$\mathbf{c(v)} - \mathbf{c(v, a)} \geq \mathbf{p_{\neg a} c(v)} + \mathbf{1},$$

gdzie

- $\mathbf{c(v, a)}$ — zliczenie wystąpień argumentu **a** z czasownikiem **v**,
- $\mathbf{c(v)}$ — zliczenie wystąpień czasownika **v**,
- $\mathbf{p_a}$ oraz $\mathbf{p_{\neg a}}$ — parametry dobrane tak, aby zminimalizować liczbę błędów dla czasowników z próby uczącej.

Czyszczenie macierzy współwystąpień

- S — relacja $\mathbf{M(v)_{ab}}$ dla nieoczyszczonej macierzy
- R — najczęstsza relacja między **a** i **b** w próbie uczącej

	zgodność z próbą testową
$\mathbf{M(v)_{ab}} := \mathrm{S}$	76%
$\mathbf{M(v)_{ab}} := \mathrm{R}$	79%
$\mathbf{M(v)_{ab}} := \mathrm{T}$	82%

- $\mathrm{T} = \begin{cases} \mathrm{R}, & \mathbf{C(a\,R\,b)} \geq \mathbf{p}_{\mathrm{S}\Rightarrow\mathrm{R}}\mathbf{C(ab)} + \mathbf{t}_{\mathrm{S}\Rightarrow\mathrm{R}}, \\ \mathrm{S}, & \text{inaczej}, \end{cases}$
- $\mathbf{C(ab)}$ — liczba czasowników o argumentach **a** i **b**
- $\mathbf{C(a\,R\,b)}$ — liczba czasowników, dla których $\mathbf{a}\,\mathrm{R}\,\mathbf{b}$
- $\mathbf{p}_{\mathrm{S}\Rightarrow\mathrm{R}}$ oraz $\mathbf{t}_{\mathrm{S}\Rightarrow\mathrm{R}}$ — parametry optymalizowane

Porównanie ze słownikami testowymi

Liczby par **(v, a)**:

a) po oczyszczeniu zbiorów argumentów:

L	AE	Bań.	Pol.	Świ.	MV
AE	680				
Bań.	592	1266			
Pol.	572	902	1243		
Świ.	571	843	891	1162	
MV	585	995	1043	984	1136
REC	0.51	0.88	0.92	0.87	
PRE	0.86	0.79	0.84	0.85	

b) po oczyszczeniu macierzy współwystąpień:

L	AE	Bań.	Pol.	Świ.	MV
AE	658				
Bań.	579	1266			
Pol.	560	902	1243		
Świ.	560	843	891	1162	
MV	573	995	1043	984	1136
REC	0.5	0.88	0.92	0.87	
PRE	0.87	0.79	0.84	0.85	

Liczby par **(v, f)**:

F	AE	Bań.	Pol.	Świ.	MV
AE	997				
Bań.	476	1571			
Pol.	393	728	1431		
Świ.	396	709	717	1245	
MV	429	927	935	916	1134
REC	0.38	0.82	0.82	0.81	
PRE	0.43	0.59	0.65	0.74	

F	AE	Bań.	Pol.	Świ.	MV
AE	658				
Bań.	408	1571			
Pol.	345	728	1431		
Świ.	350	709	717	1245	
MV	383	927	935	916	1134
REC	0.34	0.82	0.82	0.81	
PRE	0.58	0.59	0.65	0.74	

Porównanie z wcześniejszym eksperymentem

Fast i Przepiórkowski (2005):

- walencje pozyskiwane z Korpusu IPI PAN,
- płytki parser,
- metoda Brenta (1993),
- trenowanie i testowanie na słowniku Świdzińskiego,
- argumentami tylko frazy nominalne i przyimkowe.

	REC	PRE
Fast i Przepiórkowski (2005)	**48%**	**49%**
nasze wyniki (zrzutowane)		
— po oczyszczeniu zbiorów argumentów	**47%**	**58%**
— po oczyszczeniu macierzy współwystąpień	**42%**	**66%**

Niedomogi proponowanego podejścia

- Rozkład błędów w fałszywych obserwacjach pozytywnych:

FP	ogółem	argument climbing	błędy w słowniku
np(acc)	18	4	1
sie	4	0	4
np(dat)	8	1	6

- Czyszczenie macierzy współwystąpień powoduje, że wiele ich elementów staje się niezależnych od czasownika.
- Wykluczanie ZE $\times$ np(nom) jest nietypowe, a warunkowane nieobecnością *się* jest nieopisywalne:

$$\mathbf{F}(\textit{dziwić}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{ZE}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{sie}\}, \\ \{\text{np(nom)}, \text{sie}, \text{np(dat)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{ZE}\} \end{array} \right\}$$

Dziękujemy!