Automatyczne rozstrzyganie problemu PP attachment w języku polskim

Katarzyna Krasnowska

IPI PAN
Praca jest współfinansowana ze środków Unii Europejskiej
w ramach Europejskiego Funduszu Społecznego.
Projekt PO KL „Technologie Informacyjne: Badania i ich interdyscyplinarne zastosowania".

1 grudnia 2014

Problem PP attachment

Przykład:

Jan podał piłkę do drewna.

vs.

Pomocnik podał piłkę do napastnika.

Problem PP attachment

Przykład:

Jan podał piłkę do drewna.

vs.

Pomocnik podał piłkę do napastnika.

Łatwiej, mając do dyspozycji:

Problem PP attachment

Przykład:

*Jan **podał piłkę** do drewna.*

vs.

***Pomocnik podał piłkę** do **napastnika**.*

Łatwiej, mając do dyspozycji:

- ujednoznacznianie sensów słów,

Problem PP attachment

Przykład:

*Jan podał **piłkę** do **drewna**.*

vs.

*Pomocnik **podał** piłkę do **napastnika**.*

Łatwiej, mając do dyspozycji:

- ujednoznacznianie sensów słów,
- preferencje selekcyjne/kolokacje.

Problem PP attachment

Przykład ze Składnicy:

Wobec wyników głosowania stwierdzam, że Senat podjął uchwałę w sprawie ustawy$_1$ o zmianie ustawy$_2$ o organizowaniu i prowadzeniu działalności kulturalnej.

Problem PP attachment

Przykład ze Składnicy:

Wobec wyników głosowania stwierdzam, że Senat podjął uchwałę w sprawie ustawy$_1$ o zmianie ustawy$_2$ o organizowaniu i prowadzeniu działalności kulturalnej.

- ***stwierdzam*** *[wobec wyników]*
- *podjął/**uchwałę**/ustawy$_1$/ustawy$_2$/działalności/kulturalnej [w sprawie]*
- *podjął/uchwałę/sprawie/**ustawy$_1$**/ustawy$_2$/działalności/kulturalnej [o zmianie]*
- *podjął/uchwałę/sprawie/ustawy$_1$/zmianie/**ustawy$_2$**/działalności/kulturalnej [o organizowaniu i prowadzeniu]*

Problem PP attachment

Przykład ze Składnicy:

Wobec wyników głosowania stwierdzam, że Senat podjął uchwałę w sprawie ustawy$_1$ o zmianie ustawy$_2$ o organizowaniu i prowadzeniu działalności kulturalnej.

- ***stwierdzam*** *[wobec wyników]*
- *podjął/**uchwałę**/ustawy$_1$/ustawy$_2$/działalności/kulturalnej [w sprawie]*
- *podjął/uchwałę/sprawie/**ustawy$_1$**/ustawy$_2$/działalności/kulturalnej [o zmianie]*
- *podjął/uchwałę/sprawie/ustawy$_1$/zmianie/**ustawy$_2$**/działalności/kulturalnej [o organizowaniu i prowadzeniu]*
- *stwierdzam [wobec głosowania]*
- *podjął/uchwałę/działalności/kulturalnej [w prowadzeniu]*

Typowe sformułowanie problemu

Dane: czwórka (v, n, p, $n2$):

- v – potencjalny nadrzędnik czasownikowy,
- n – potencjalny nadrzędnik rzeczownikowy,
- p, $n2$ – główne elementy frazy przyimkowej.

Wynik: decyzja binarna V/N, np.:

- *Jan podał piłkę do drewna.*
  $\rightarrow$ (*podać*, *piłka*, *do:gen*, *drewno*) $\rightarrow N$
- *Pomocnik podał piłkę do napastnika.*
  $\rightarrow$ (*podać*, *piłka*, *do:gen*, *napastnik*) $\rightarrow V$

Dane – uczenie z nadzorem

Krzaki (`http://zil.ipipan.waw.pl/Krzaki`):

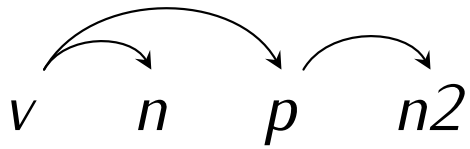
- ręcznie stworzone drzewa zależnościowe,
- zestaw 20000 zdań wylosowanych do Składnicy.

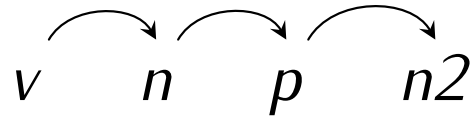
5734 piątki (v, n, p, $n2$, a):

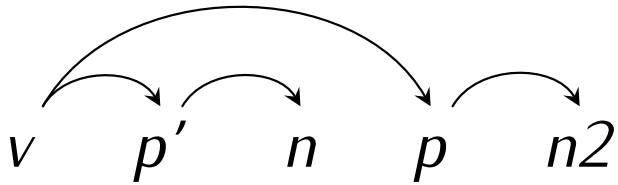
- n jest podrzędnikiem v,
- $n2$ jest podrzędnikiem p,
- p jest podrzędnikiem v (a=V) lub n (a=N).

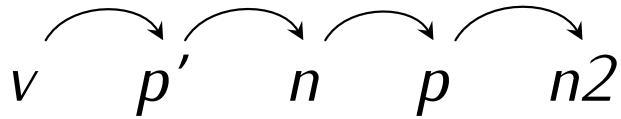
	KRZAKI-TRAIN	KRZAKI-DEV	KRZAKI-TEST
BY-TUPLE	2867 (50%)	1434 (25%)	1433 (25%)
BY-SENTENCE	2810 (49%)	1504 (26%)	1420 (25%)

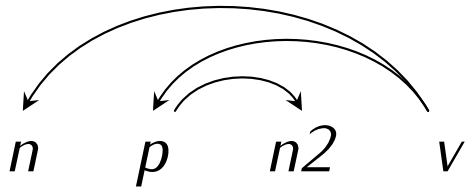
Dane – uczenie z nadzorem

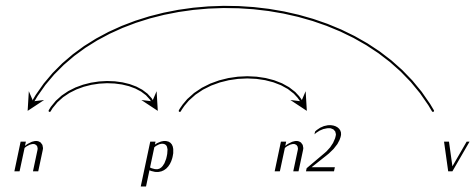


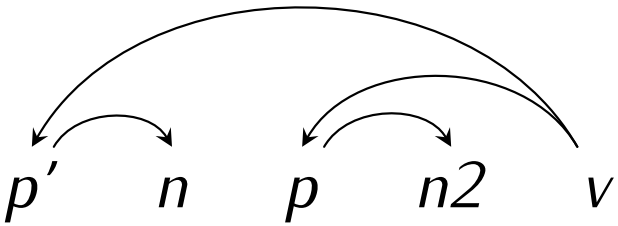


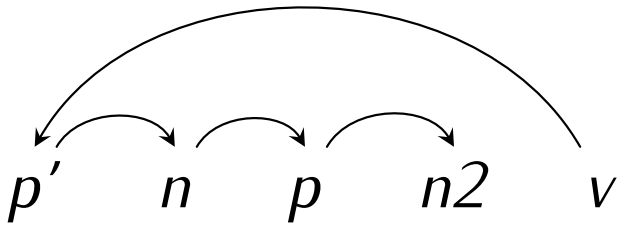












$(v, n, p, n2, \mathbf{V})$ $(v, n, p, n2, \mathbf{N})$

Dane – uczenie z częściowym nadzorem

Pomysł: wykorzystanie dużych, ale „niepełnych” danych z NKJP.

- `[pos="praet"] [pos="subst"] [pos="prep"] [pos="subst"]`:
  - „wyłapali błędy w przekładach”,
  - „prosiła znajomych o obronę”,
  - „odniósł triumf nad wrogiem”.
- $\longrightarrow$ niejednoznaczne dane: (v, n, p, $n2$, ?).
- Zamiast tego:
  - ok. 18 000 000 trójek (v, p, $n2$),
  - ok. 3 000 000 trójek (n, p, $n2$).

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*v*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="fin|praet|impt|imps|inf|pcon|pact|ger"]
  [pos="qub" & base="by"]? [pos="aglt"]?
  [pos="qub" & base="się"]? [pos="adv"]{,2}
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...) się całkiem niedawno przeprowadził do bardzo dużego domu (...)
- (...) zdziwiłabym się ogromnie na widok (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
- (...), że najlepsze gofry z bitą śmietaną (...)

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

Zapytanie dla (*n*, *p*, *n*2) (pełny NKJP):

- ```
  [pos="interp"] [pos="conj|comp"]?
  [pos="adv"]{,2}
  [pos="adj" & case~$2]* [pos="subst" & case~$2]
  [pos="prep" & case~$1] [pos="adv"]{,2}
  [pos="adj" & case~$1]* [pos="subst" & case~$1]
  meta channel="prasa|ksiazka"
  ```
- (...), że najlepsze gofry z bitą śmietaną (...)
  ```

Dane – uczenie z częściowym nadzorem

- $c(x, p)$: liczba trójek $(x, p, *)$.
- $c(x)$: liczba trójek $(x, *, *)$.
- $c(V, p)$: liczba trójek $(v, p, *)$, gdzie v to czasownik,
- $c(N, p)$: liczba trójek $(n, p, *)$, gdzie n to rzeczownik,
- $c(V)$: liczba trójek $(v, *, *)$, gdzie v to czasownik,
- $c(N)$: liczba trójek $(n, *, *)$, gdzie n to rzeczownik.

Baseline

Trzy proste modele:

- **Zawsze** V.
- **Z częściowym nadzorem**: V wtw. $\frac{c(V,p)}{c(V)} \geqslant \frac{c(N,p)}{c(N)}$.
- **Z nadzorem** (Collins i Brooks, 1995): Przyłączenie częściej występujące dla p w danych uczących (Krzaki-train).

	BY-TUPLE	BY-SENTENCE
Zawsze V	55,4%	57,0%
Z częściowym nadzorem	54,9%	56,3%
Z nadzorem	**61,7%**	**63,4%**

Baseline

Trzy proste modele:

- **Zawsze** V.
- **Z częściowym nadzorem**: V wtw. $\frac{c(V,p)}{c(V)} \geqslant \frac{c(N,p)}{c(N)}$.
- **Z nadzorem** (Collins i Brooks, 1995): Przyłączenie częściej występujące dla p w danych uczących (Krzaki-train).

	BY-TUPLE	BY-SENTENCE
Zawsze V	55,4%	57,0%
Z częściowym nadzorem	54,9%	56,3%
Z nadzorem	**61,7%**	**63,4%**

Dla angielskiego:

- **Zawsze** N: 67%, 59%,
- **Z nadzorem**: 72,2%.

Metoda z częściowym nadzorem

Prawdopodobieństwa szacowane na podstawie danych z NKJP (wg Hindle i Rooth 1993):

- $P(p|v) = \frac{c(v,p)+\frac{c(V,p)}{c(V)}}{c(v)+1}$,
- $P(p|n) = \frac{c(n,p)+\frac{c(N,p)}{c(N)}}{c(n)+1}$.

Nadrzędnik $= V$ wtw. $\log P(p|v) - \log P(p|n) \geqslant -0,1$.

BY-TUPLE	BY-SENTENCE
73,3%	73,5%

Redukcja błędu względem baseline'u: 30,3% i 27,6%.

Kategorie semantyczne w Słowosieci

Przykłady:

lemat	znaczenie	kategoria
chleb	1	jedzenie
kot	1, 2, 6	zwierzęta
	3, 4	ludzie
	5	wytwory ludzkie
sąd	1	grupy ludzi i rzeczy
	2	związek miedzy ludźmi, rzeczami lub ideami
	3	miejsca i umiejscowienie
	4	zdarzenia
	5, 6	związane z myśleniem

Metoda z częściowym nadzorem (wersja II)

- Każda trójka (h, p, $n2$) z NKJP traktowana jako k trójek (h, $p + c_i$, $n2$), każda o wadze $\frac{1}{k}$, gdzie:
  - k – liczba znaczeń $n2$ w Słowosieci,
  - c_i – kategoria semantyczna i-tego znaczenia.
- Klasyfikacja dla czwórki (n, v, p, $n2$):
  - wybór V/N osobno dla każdej czwórki (v, n, $p + c_i$, $n2$),
  - ostateczna decyzja: „głosowanie”.

BY-TUPLE	BY-SENTENCE
76,9%	75,5%

Redukcja błędu względem wersji podstawowej: 13,5% i 7,5%.

Metoda z nadzorem: backed-off model

Collins i Brooks (1995): dla czwórki $(v, n, p, n2)$ wybierane jest najczęstsze podłączenie wśród podobnych danych uczących:

poziom	postać podobnych krotek
4	$(v, n, p, n2, a)$
3	$(v', n, p, n2, a)$, $(v, n', p, n2, a)$, $(v, n, p, n2', a)$
2	$(v', n', p, n2, a)$, $(v', n, p, n2', a)$, $(v, n', p, n2', a)$
1	$(v', n', p, n2', a)$
0	domyślnie: V

Wyniki

d. uczące	d. testowe	4	3	2	1	0	
KRZAKI-TRAIN BY-TUPLE	KRZAKI-DEV BY-TUPLE	94,7%	95,4%	72,8%	59,7%	50,0%	**73,2%**
	sklasyfikowanych	19	260	717	434	4	
	pokrycie	1,3%	19,5%	69,5%	99,7%	100%	
KRZAKI-TRAIN BY-SENTENCE	KRZAKI-DEV BY-SENTENCE	86,7%	83,5%	72,2%	66,3%	76,2%	**71,3%**
	sklasyfikowanych	15	103	893	472	21	
	pokrycie	1.0%	7,8%	67,2%	98,6%	100%	

Wyniki

d. uczące	d. testowe	4	3	2	1	0	
Projected	Krzaki-dev by-tuple	82,5%	66,4%	65,3%	63,2%	100,0%	**66,3%**
	sklasyfikowanych	57	458	880	38	1	
	pokrycie	4.0%	35,9%	97,3%	99,9%	100%	
Projected	Krzaki-dev by-sentence	72,4%	70,9%	66,4%	61,5%	100,0%	**68.0%**
	sklasyfikowanych	58	506	887	52	1	
	pokrycie	3,9%	37,5%	96,5%	99,9%	100%	

Wyniki

d. uczące	d. testowe	4	3	2	1	0	
Krzaki-CV by-tuple	Krzaki-CV by-tuple	90,6%	93,4%	73,6%	64,3%	57,1%	**76,1%**
	sklasyfikowanych	117	1205	3197	1201	14	
	pokrycie	2.0%	23,1%	78,8%	99,8%	100,0%	
Krzaki-CV by-sentence	Krzaki-CV by-sentence	89,6%	85,6%	74,9%	64,9%	66,7%	**73,8%**
	sklasyfikowanych	67	618	3660	1371	18	
	pokrycie	1,2%	11,9%	75,8%	99,7%	100,0%	
Projected	Krzaki by-tuple	73,7%	69,3%	65,3%	58,6%	88,9%	**66,8%**
	sklasyfikowanych	232	1940	3384	169	9	
	pokrycie	4.0%	37,9%	96,9%	99,8%	100,0%	

Rozszerzenie I: synsety Słowosieci

- Dla czwórki $(v, n, p, n2)$ generowane są wszystkie czwórki $(v, n', p, n2')$, takie, że:
  - n' jest w synsecie n,
  - $n2'$ jest w synsecie $n2$.
- Klasyfikacja dla każdej krotki osobno.
- Ostateczna decyzja: częstsza na najwyższym występującym poziomie.

d. uczące	d. testowe	4	3	2	1	0	
KRZAKI-TRAIN BY-SENTENCE	KRZAKI-DEV BY-SENTENCE	87,5%	80,7%	71,3%	65,9%	76,2%	**70,8%**
	sklasyfikowanych	16	119	931	417	21	
	pokrycie	1,1%	9.0%	70,9%	98,6%	100%	

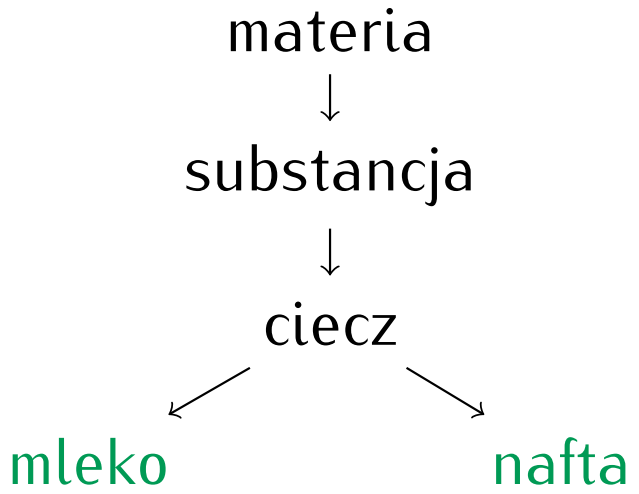
Rozszerzenie II: listy podobieństwa

krowa				
bydło	owca	świnia	koza	wół
zwierzę	jałówka	koń	cielę	osioł
wielbłąd	królik	pies	jagnię	drób
bawół	kura	byk	kot	szczur
wiolonczelista				
barakowóz	marlin	dyrygent	organista	pianista
skrzypek	powijak	muzykolog	ogryzek	pianistyka
reperacja	kirka	kompozytor	saksofonista	aranżer
wikipedystka	futurysta	strukturalizacja	widzewiak	filler
zamek				
pałac	twierdza	warownia	forteca	ratusz
dwór	budowla	rezydencja	wieża	kłódka
gród	kaplica	brama	zasuwa	rygiel
kościół	baszta	fortyfikacja	drzwi	pałacyk

Rozszerzenie II: listy podobieństwa

d. uczące	d. testowe	4	3	2	1	0	
KRZAKI-TRAIN BY-SENTENCE	KRZAKI-DEV BY-SENTENCE	78,9%	80,7%	70,9%	68,1%	76,2%	**72,3%**
	sklasyfikowanych	19	259	995	210	21	
	pokrycie	1,3%	18,5%	84,6%	98,6%	100,0%	

Odległość wordnetowa



Odległość wordnetowa (Wu i Palmer, 1994)

$$dist(s_1, s_2) = 1 - \frac{2 \cdot depth(H)}{d(s_1, H) + d(s_2, H) + 2 \cdot depth(H)}$$

- H – najniższy wspólny hiperonim s_1 i s_2.
- $d(s, s')$ – długość ścieżki łączącej s z s'.
- Intuicja: „drobnoziarnistość" podziału pojęć rośnie wraz z głębokością.
- Dla poprzedniego przykładu:
  - $dist(lew, tygrys) = 1 - \frac{2 \cdot 6}{1+1+2 \cdot 6} = 0,14$
  - $dist(mleko, nafta) = 1 - \frac{2 \cdot 2}{1+1+2 \cdot 2} = 0,33$
- Odległość między słowami w_1 i w_2:

$$dist(w_1, w_2) = \min_{s_1 \ni w_1, s_2 \ni w_2} dist(s_1, s_2).$$

Modyfikacja relacji hiperonimii w Słowosieci

- Miara Wu-Palmer operuje na strukturze hiperonimii.
- Nie każdy synset w Słowosieci posiada hiperonimy.

relacja	izolowane synsety	%
tylko hiperonimia	13 814	12,4%
+ bliskoznaczność	10 861	9,7%
+ typ/przykład	8 222	7,4%
+ nazwa żeńska	6 103	5,5%
+ nacechowanie	2 765	2,5%
+ holonimia/meronimia	241	0,2%

Rozszerzenie III: odległość wordnetowa

Czwórka do klasyfikacji: $(v, n, p, n2)$

poziom	postać podobnych krotek
4	$(v', n', p, n2', a)$
3	$(v'', n', p, n2', a)$,
	$(v', n'', p, n2', a)$,
	$(v', n', p, n2'', a)$
2	$(v'', n'', p, n2', a)$,
	$(v'', n', p, n2'', a)$,
	$(v', n'', p, n2'', a)$
1	$(v'', n'', p, n2'', a)$
0	domyślnie: V

$$dist(v, v') \leqslant t_v, \quad dist(n, n') \leqslant t_n, \quad dist(n2, n2') \leqslant t_{n2}$$

Rozszerzenie III: odległość wordnetowa

WSD	t_v	t_n	t_{n2}	
−	0,0	0,0	0,0	72,0%
−	0,05	0,05	0,05	71,9%
−	0,1	0,1	0,1	72,1%
−	0,15	0,15	0,15	71,5%
+	0,0	0,0	0,0	72,5%
+	0,05	0,05	0,05	72,5%
+	0,1	0,1	0,1	72,0%
+	0,15	0,15	0,15	71,7%
+	0,0	0,0	0,1	72,7%
+	0,0	0,0	0,15	72,7%
+	0,05	0,05	0,15	72,7%

Backed-off model: podsumowanie

wariant	pokrycie na poz. 3-4	wynik
standardowy	8,8%	71,3%
synsety	10,1%	70,8%
listy podobieństwa	19,8%	72,3%
odległość wordnetowa (WSD, t_v, t_n=0,05, t_{n2}=0,15)	12.3%	72,7 %

Proste połączenie metod

- Decyzja modelu z nadzorem (wersja z listami podobieństwa), jeśli poziom $\geqslant 3$.
- W przeciwnym przypadku decyzja modelu z częściowym nadzorem (wersja z kategoriami semantycznymi).

model	Krzaki-dev	Krzaki-test
z częściowym nadzorem	75,5%	75,7%
z nadzorem	72,3%	69,6%
połączenie	76,9%	75,3%

Porównanie z działaniem człowieka

Eksperyment pilotażowy (1 anotator):

- Próbka 200 czwórek ($v, n, p, n2$).
- Możliwe decyzje: V, N, „?”.
- Zgodność z Krzakami:
  - 79% dla całej próbki,
  - 83% po odrzuceniu „?”.

Ręcznie anotowana wersja danych z Krzaków:

- 2 anotatorów + superanotator,
- 78,3% potwierdzonych bez interwencji superanotatora,
- 80,6% zaakceptowanych bez interwencji superanotatora,
- 8,6% odrzuconych.

Dalsze eksperymenty

- Inne metody uczenia.
- Szerszy repertuar cech:
  - wyniki opisanych metod,
  - cechy morfoskładniowe,
  - informacja walencyjna,
  - szerszy kontekst.
- Połączenie z pełnym ujednoznacznianiem składniowym

Dalsze eksperymenty

- Inne metody uczenia.
- Szerszy repertuar cech:
  - wyniki opisanych metod,
  - cechy morfoskładniowe,
  - informacja walencyjna,
  - szerszy kontekst.
- Połączenie z pełnym ujednoznacznianiem składniowym

Dziękuję za uwagę.

Bibliografia

Collins, M. i Brooks, J. (1995). Prepositional Phrase Attachment through a Backed-Off Model. W: *Proceedings of the Third Workshop on Very Large Corpora*, str. 27–38.

Hindle, D. i Rooth, M. (1993). Structural Ambiguity and Lexical Relations. *Computational Linguistics*, **19**(1), 103–120.

Wu, Z. i Palmer, M. (1994). Verbs Semantics and Lexical Selection. W: *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, str. 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.