

Metody automatycznego wykrywania błędów w bankach drzew

Katarzyna Krasnowska

IPI PAN

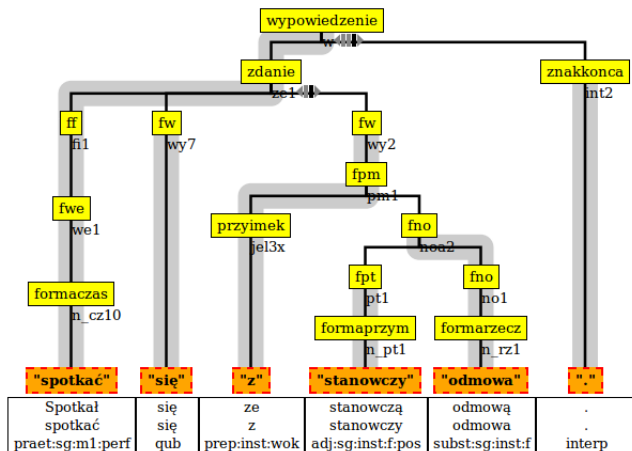
26 listopada 2012

Plan prezentacji

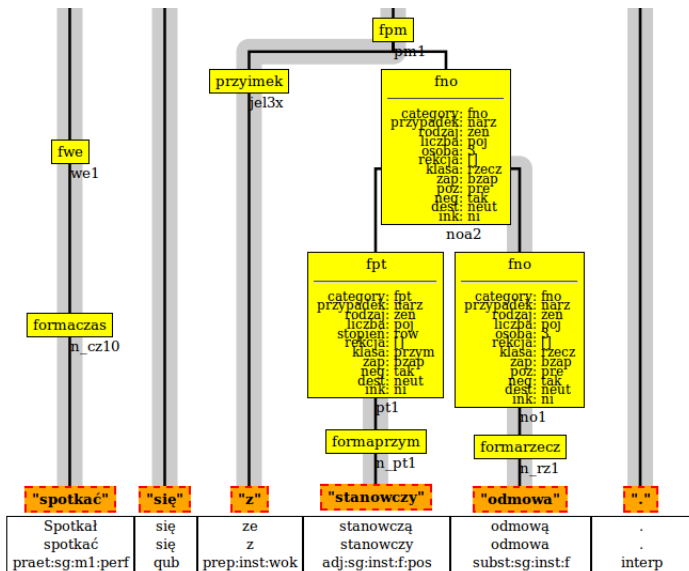
- 1 *Składnica – pierwsze próby*
- 2 *Składnica – metoda STSG*
- 3 *Składnica zależnościowa*
- 4 *Treebank rzutowany*

- Polski bank drzew (Świdziński i Woliński, 2010)
- Aktualna wersja zawiera 8227 drzew składniowych
- Drzewa budowane półautomatycznie
- Węzły wewnętrzne posiadają kategorię składniową oraz zestaw atrybutów

Składnica – przykład drzewa



Składnica – przykład drzewa



Składnica vs Penn Treebank (WSJ)

- Rozmiar korpusu: 8227 vs 49208 drzew
- Sposób konstrukcji: półautomatyczny vs ręczny
- Poziom zróżnicowania tekstów:
 - *Składnica*: literatura, prasa, obrady Sejmu ...
 - WSJ: dziennik Wall Street Journal
- Fleksyjność języka

(Dickinson i Meurers, 2005):

- Immediate Dominance Sets – zbiory bezpośredniej dominacji
- Dla każdego występującego w treebanku rodzeństwa wyznaczany jest zbiór wszystkich ojców
- Zróżnicowanie takiego zbioru sugeruje błąd w anotacji
 - np. rodzeństwo $\langle \text{ADJP}, \text{NP} \rangle$ występuje raz jako dzieci wężła NP, a raz — wężła ADJP
- Wyniki na WSJ:
 - poprawność (precision) 65,6%
 - pełność (recall) 43,0%

Zbiory ID – wyniki eksperymentów

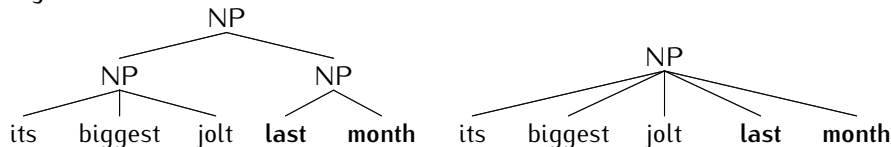
- W pierwszym podejściu: 1254 (16,5%) drzew oznaczonych jako błędne
 - niemal wyłącznie „falszywe alarmy” związane z etykietami fw/fl
- Po wyeliminowaniu węzłów fw/fl:
 - 103 drzew wskazane jako błędne
 - w tym ~80% niestusznie

Metoda n -gramowa

(Dickinson i Meurers, 2003):

- W banku drzew wyszukiwane są wystąpienia identycznych ciągów słów, dla których:
 - węzły dominujące mają różne etykiety ...
 - ... lub tylko jeden z nich posiada węzeł dominujący (tj. drugi nie został opisany jako fraza)
- Im dłuższy wspólny kontekst, tym lepiej

Przykład:



Kontekst: *market received its biggest jolt last month from Campeau Corp., which...*

Metoda n -gramowa – wyniki eksperymentów

n -gramy form ortograficznych bez kontekstu (etykieta = kategoria składniowa + pełen zestaw atrybutów):

długość ciągu	1	2	3	4	5
liczba ciągów	4375	881	108	14	3

- Dickinson i Meurers (2003): maksymalna długość 46
- Dla długości 1 niemal wyłącznie niejednoznaczności typu
*widzę **kobiety** vs **kobiety przyszły***
- Przy wymaganiu obustronnego kontekstu długości 1 liczba n -gramów spadła do 56

Metoda n -gramowa – wyniki eksperymentów

n -gramy form ortograficznych bez kontekstu (etykieta = kategoria składniowa + pełen zestaw atrybutów):

długość	n -gramy			zdania		
	sprawdzone	prawidłowe		sprawdzone	prawidłowe	
2	31	5	16,1%	130	12	9,2%
3	22	3	13,6%	59	7	11,9%
4	14	8	57,1%	32	19	59,4%
5	3	3	100,0%	6	6	100,0%
razem	70	19	27,1%	227	44	19,4%

Metoda n -gramowa – wyniki eksperymentów

n -gramy tagów morfoskładniowych bez kontekstu:

długość n -gramu	2	3	4	5	6	7
liczba n -gramów	1516	1760	599	91	9	2

n -gramy tagów morfoskładniowych z kontekstem:

min. długość kontekstu	2	3	4	5	6
liczba n -gramów	5511	499	88	35	16

Metoda n -gramowa – wyniki eksperymentów

min. długość kontekstu	2	3
rozmiar próbki	52	39
poprawność (precision)	23,1%	44,8%

Przykłady n -gramów

Kontekst długości 2:

Czy ktoś jeszcze_{qub} [chciał_{praet} by_{qub}]formaczas się_{qub} zapisać?

Nie_{qub} [orientował_{praet} się_{qub}]NIL też_{qub}, że popełnia przestępstwo (...).

— „fałszywy alarm” — oba zdania są prawidłowo opisane

kontekst długości 3:

Minęto_{praet} [już_{qub} kilka_{num}]ficz dni_{subst:pl:gen:m3} ,interp odkąd (...).

Do obozu pozostało_{praet} [jeszcze_{qub} parę_{num}]NIL kilometrów_{subst} ·interp

— błąd w Składnicy

Przykłady n -gramów

kontekst długości 4:

Daktyle (...) dodać z_{prep} [*pozostałym* $_{\text{adj}}$ *masłem* $_{\text{subst}}$] $_{\text{fno}}$ do_{prep} *kuskusu* $_{\text{subst}}$

\cdot interp

Z_{prep} [*oficjalnym* $_{\text{adj}}$ *żądaniem* $_{\text{subst}}$] $_{\text{NIL}}$ *wobec* $_{\text{prep}}$ *MKOl* $_{\text{subst}}$ \cdot interp (...) *wystąpił*
Narodowy Komitet Olimpijski KRLD (...).

— „fałszywy alarm”

kontekst długości 6:

Złotówka straciła też $_{\text{qub}}$ [*na* $_{\text{prep}}$ *wartości* $_{\text{subst}}$] $_{\text{fpm}}$ w_{prep} *stosunku* $_{\text{subst}}$ do_{prep}

euro $_{\text{subst:sg:gen:n}}$ \cdot interp

(...) *uczniowie ustawiają się* $_{\text{qub}}$ [*na* $_{\text{prep}}$ *zbiórce* $_{\text{subst}}$] $_{\text{NIL}}$ w_{prep} *parku* $_{\text{subst}}$
naprzeciwko $_{\text{prep}}$ *gimnazjum* $_{\text{subst}}$ \cdot interp

— „fałszywy alarm”

Plan prezentacji

- 1 *Składnica – pierwsze próby*
- 2 *Składnica – metoda STSG*
- 3 *Składnica zależnościowa*
- 4 *Treebank rzutowany*

Synchronous Tree Substitution Grammar

Cohn i Lapata (2009), Kato i Matsubara (2010):

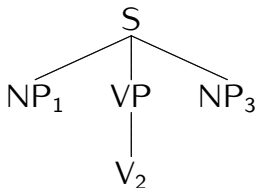
- Każda reguła składa się z:
 - pary drzew elementarnych (**elementary trees**)
 - wyrównania 1-do-1 (**alignment**) między ich liśćmi
- Przykład reguły:

Synchronous Tree Substitution Grammar

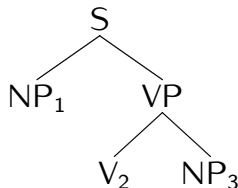
Cohn i Lapata (2009), Kato i Matsubara (2010):

- Każda reguła składa się z:
 - pary drzew elementarnych (**elementary trees**)
 - wyrównania 1-do-1 (**alignment**) między ich liśćmi
- Przykład reguły:

źródło:

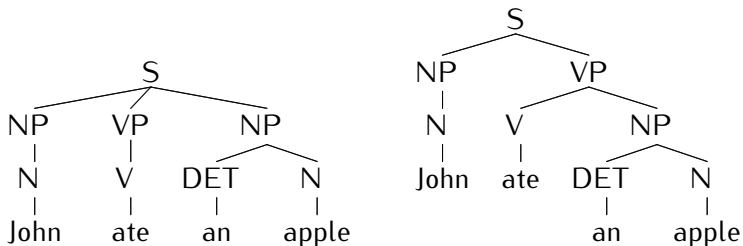


cel:



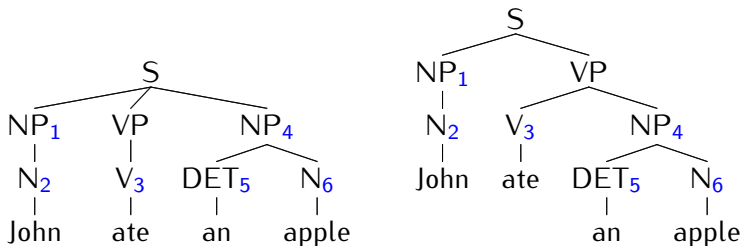
Uzyskiwanie reguł STSG z banku drzew

Dla każdych dwóch poddrzew o identycznych etykietach korzeni i planach...



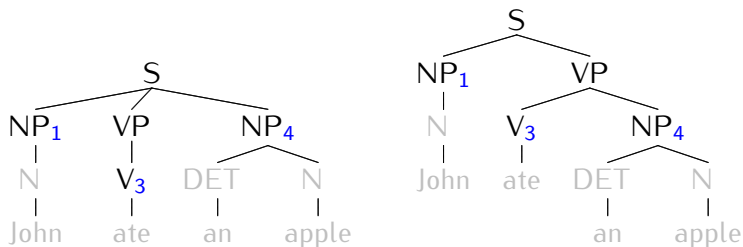
Uzyskiwanie reguł STSG z banku drzew

...dokonywane jest wyrównanie pomiędzy ich węzłami wewnętrznymi o identycznych etykietach korzeni i płonach.



Uzyskiwanie reguł STSG z banku drzew

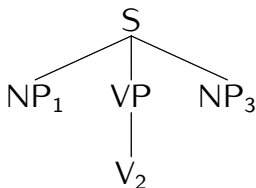
Dzieci każdego wyrównanego węzła są następnie usuwane...



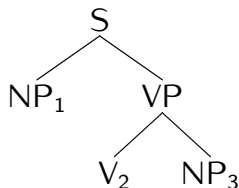
Uzyskiwanie reguł STSG z banku drzew

...w rezultacie powstaje reguła:

źródło:



cel:



Wybór reguł STSG

- Dla każdej powstałej w ten sposób reguły, powstanie również reguła „odwrotna”
- Reguły o poprawnym źródle i niepoprawnym celu powinny zostać wyeliminowane
- Spodziewamy się, że poprawne konstrukcje są częstsze od błędnych
- $f(\tau)$ – częstość występowania drzewa elementarnego τ w banku, $0 \leq f(\tau) \leq 1$
- Wybierane są reguły o wysokiej wartości *Score*:

$$\text{Score}(\langle \tau, \tau' \rangle) = \frac{f(\tau')}{f(\tau) + f(\tau')}$$

Eksperyment na WSJ

- 49 208 zdań
- 8 776 reguł
- Użyto 100 najlepszych wg. miary *Score*:

Eksperyment na WSJ

- 49 208 zdań
- 8 776 reguł
- Użyto 100 najlepszych wg. miary *Score*:
 - 331 zmian w korpusie
 - dokładność: 71,9%
 - dla 70 reguł dokładność wyniosła 100%
 - 30 reguł zmieniających strukturę drzew

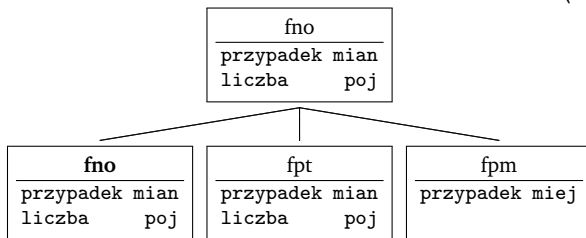
Modyfikacje metody

(Krasnowska *et al.*, 2012)

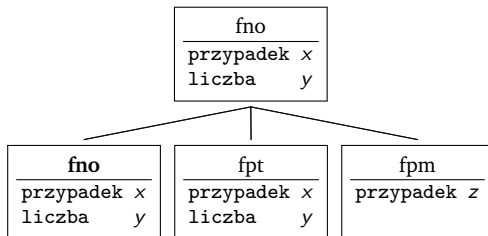
- Tagi morfoskładniowe zamiast form ortograficznych
 - możliwość brania pod uwagę formy bazowej dla kublików, przyimków, spójników i predykatywów
 - możliwość wyboru składników tagów morfoskładniowych
- Kategoria składniowa oraz wybrane atrybuty jako etykiety
- Zastąpienie wartości wybranych atrybutów zmiennymi

„Uzmiennianie” atrybutów – przykład

(1)

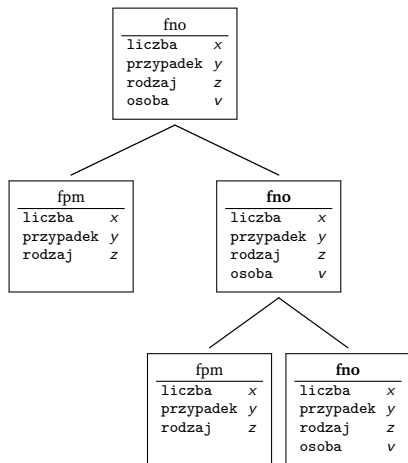


(2)

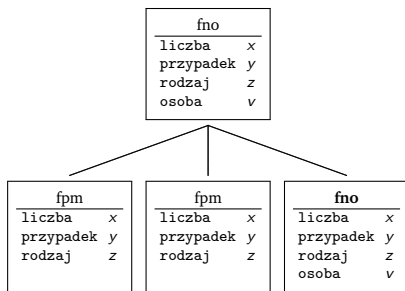


Przykładowa reguła

źródło:



cel:



Wyniki eksperymentu na *Składnicy*

Typ wyniku	Komentarz
ERR_0	Wykryty błąd, prawidłowa propozycja poprawy.
INC_0	Wykryta niespójność, prawidłowa propozycja poprawy.
ERR_1	Wykryty błąd, nieprawidłowa propozycja poprawy.
INC_1	Wykryta niespójność, nieprawidłowa propozycja poprawy.
FP	Fatszywy alarm.

Wyniki eksperymentu na *Składnicy*

$$P_0 = \frac{ERR_0 + INC_0}{ALL}$$

$$P_1 = \frac{ERR_0 + INC_0 + ERR_1 + INC_1}{ALL}$$

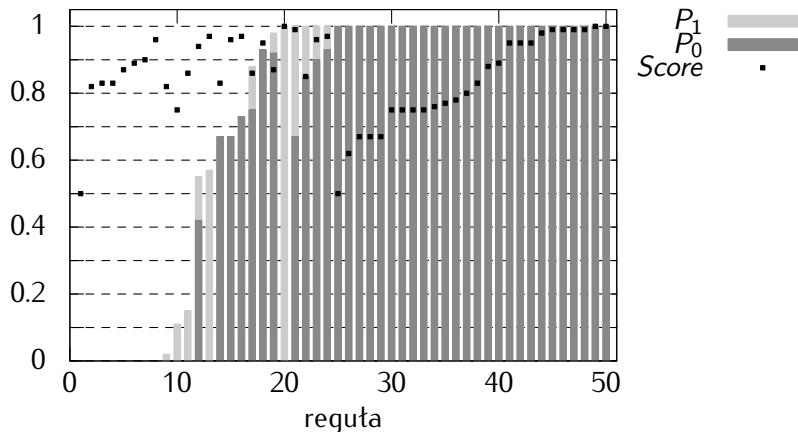
$$P_{err} = \frac{ERR_0 + ERR_1}{ALL}$$

$$R_{\sim} = \frac{ERR_0 + INC_0 + ERR_1 + INC_1}{1481}$$

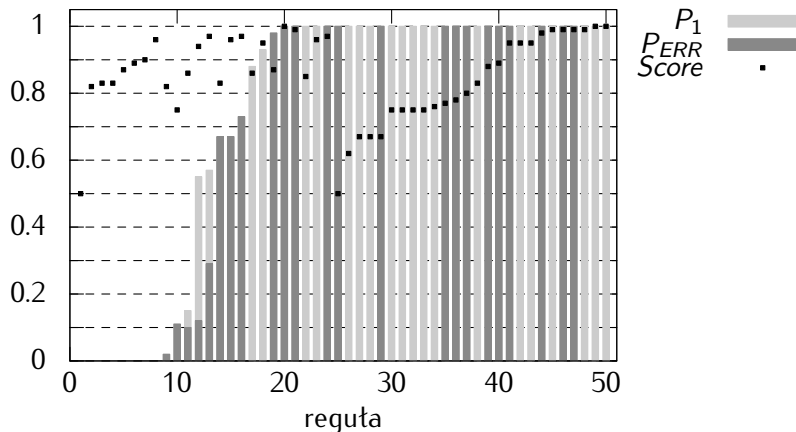
Wyniki eksperymentu na *Składnicy*

min. <i>Score</i>	P_0	P_1	R_{\sim}	F
0,5	50,33%	56,26%	17,29%	26,45%
0,6	50,33%	56,29%	17,22%	26,37%
0,7	49,44%	55,51%	16,68%	25,65%
0,8	48,82%	54,98%	15,67%	24,38%
0,9	68,04%	75,26%	9,86%	17,43%

Wyniki eksperymentu na *Składnicy*

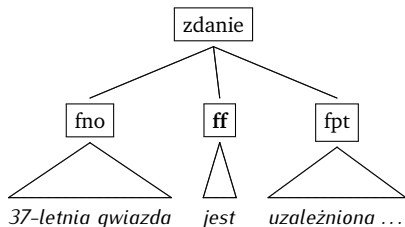
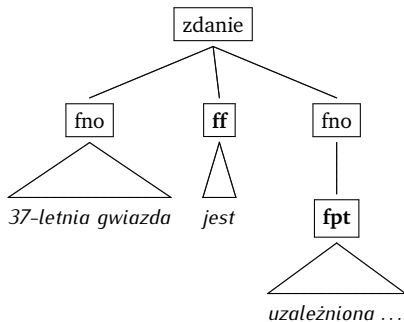


Wyniki eksperymentu na *Składnicy*



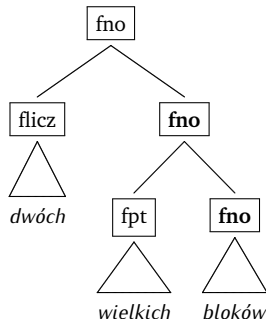
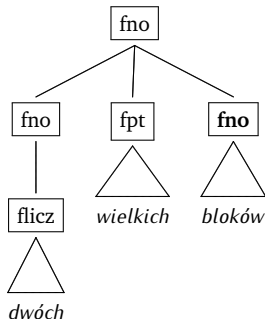
Wyniki eksperymentu na *Składnicy* – przykład

„Tymczasem wiadomo, że 37-letnia gwiazda jest uzależniona od alkoholu i narkotyków.”



Wyniki eksperymentu na *Składnicy* – przykład

„Nasz system polityczny ewoluuje w kierunku dwóch wielkich bloków.”



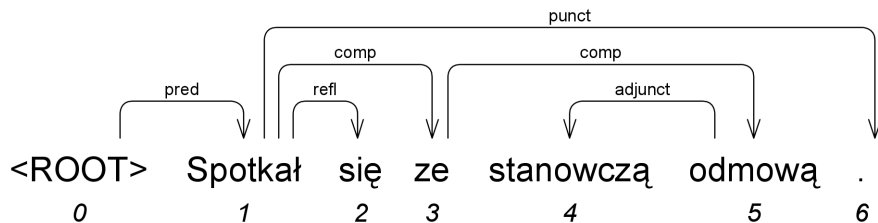
Plan prezentacji

- 1 *Składnica – pierwsze próby*
- 2 *Składnica – metoda STSG*
- 3 *Składnica zależnościowa*
- 4 Treebank rzutowany

Składnica zależnościowa

- Bank drzew zależnościowych
- Powstały w wyniku konwersji *Składnicy* (Wróblewska, 2012)
- 8227 zdań
- Motywacja: być może w drzewach w postaci zależnościowej uda się wykryć błędy anotacji, których nie wychwyciła poprzednia metoda

Składnica zależnościowa – przykład drzewa

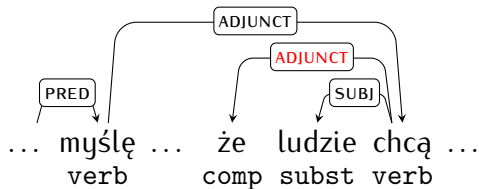


Krótko o metodzie

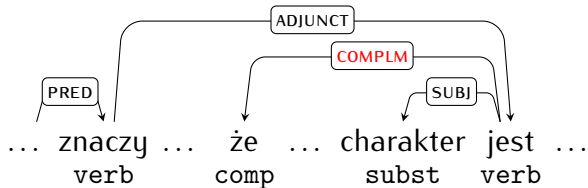
- Pary spójnych podgrafów drzew zależnościowych
- Reguła jest tworzona, jeżeli:
 - zawierają ten sam ciąg części mowy (CPOS)
 - ale różnią się strukturą (kształtem lub relacjami zależnościowymi)
- Analogiczna miara *Score* dla reguł

Przykład

„- *Ja myślę, że ludzie chcą zaoszczędzić.*”

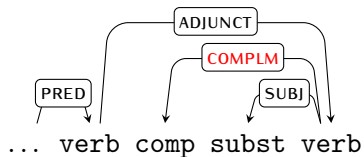
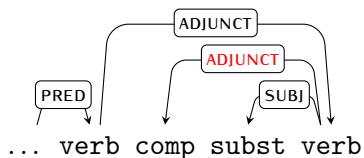


„*Jeśli to jest absolutnie niemożliwe, znaczy to, że twój charakter jest czymś pośrednim pomiędzy dwoma opisanymi w rozwiązaniu przy odpowiednich numerach.*”



Przykład – c.d.

Uzyskana reguła:



- 1000 drzew jako dane robocze
- Pozostałe 7227 drzew jako ostateczne dane testowe
- Mało obiecujące wyniki pomimo licznych prób ich poprawy:
 - dla niektóre części mowy bardziej szczegółowo rozróżniane (INF, IMPS, PRED, GER, BEDZIE)
 - dodanie do części mowy informacji o przypadku
 - forma bazowa dla kublików, interpunkcji, przyimków itp.
 - pominięcie relacji zależnościowych związanych z interpunkcją
 - wymuszenie włączania wybranych relacji do podgrafów (COMPLM, NEG, REFL)

Wyniki – 100 pierwszych zgłoszonych błędów ¹

ocena	błędy	zdania
rzeczywisty błąd, dobrze poprawiony	20	17
rzeczywisty błąd, źle poprawiony	11	11
„fałszywy alarm”	77	73
razem	108	101

¹W 99 zdaniach; niektóre reguły wprowadzały dwie różne zmiany (stąd liczba zgłoszonych błędów – 108); w dwóch takich przypadkach jedna zmiana była poprawna, a druga nie (dlatego liczby w kol. „zdania” sumują się do 101).

Wnioski z eksperymentu

- Powodem wielu niepoprawnych zgłoszeń błędu był brak informacji leksykalnej w regułach
 - na przykład dla czasownika „mieć” próba zamiany relacji COMP na OBJ
- *Składnica zależnościowa* zawiera zbyt mało zdań, aby można było tę informację wykorzystać
- Potrzebny dużo większy zasób

Plan prezentacji

- 1 *Składnica – pierwsze próby*
- 2 *Składnica – metoda STSG*
- 3 *Składnica zależnościowa*
- 4 **Treebank rzutowany**

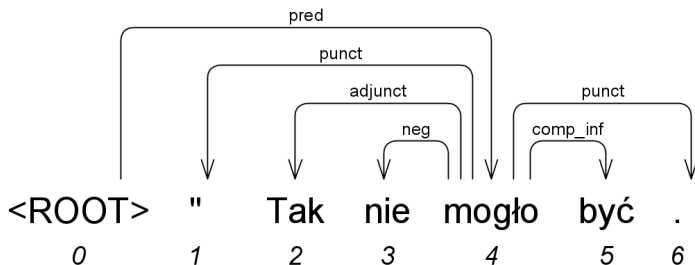
Treebank rzutowany

(Wróblewska i Przepiórkowski, 2012)

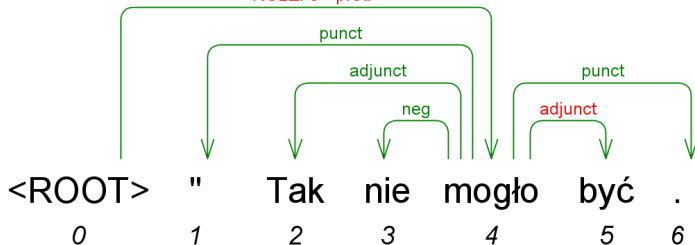
- Ponad 3 miliony zdań
- Zdania z korpusu równoległego polsko-angielskiego
- Tekst angielski sparsowany za pomocą dobrej jakości narzędzi
- Drzewa zależnościowe dla zdań polskich stworzone na podstawie zrównoleglenia

- Rozmiar zasobu pozwolił na branie pod uwagę pełnej informacji leksykalnej
 - formy bazowe słów + przypadek
- 4111 reguł
- 18885 zdań (0,6% ze wszystkich)
- Spośród 119 ręcznie sprawdzonych:
 - 57 (50%) poprawnych
 - 19 (17%) niepoprawnych (również ważna informacja!)
 - 37 (33%) nieskutecznych
- Dokładna ewaluacja trudna ze względu na rozmiar danych

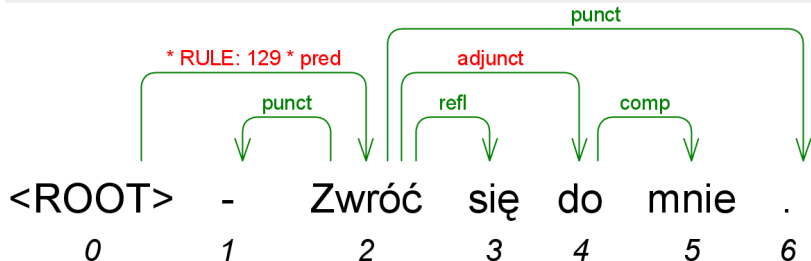
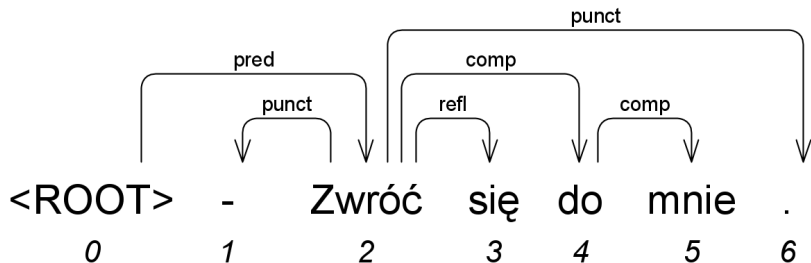
Przykład zastosowania reguły



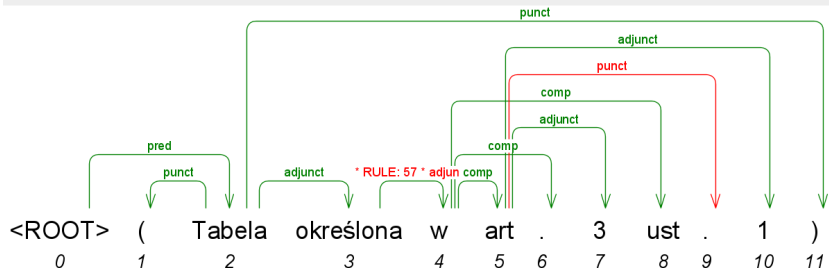
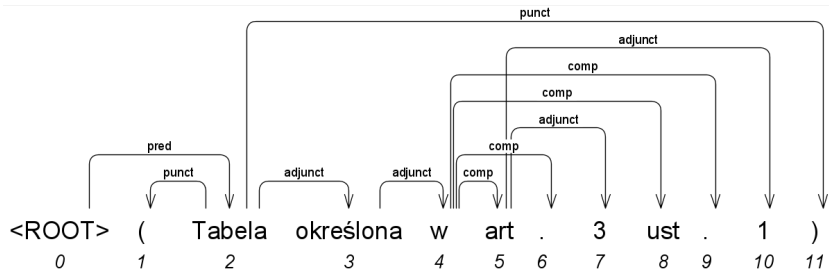
* RULE: 3 * pred



Przykład zastosowania reguły



Przykład nieskutecznej reguły



Co dalej?

- Różne pomysły na modyfikacje:
 - tylko niektóre części mowy w formie bazowej
 - więcej cech morfoskładniowych
 - „uzmiennienie” reguł
 - ...
- Problem: czasochłonna ocena skuteczności!

Co dalej?

- Pomysł: system wspomagający ewaluację metody
 - w tej chwili: *MaltEval* – tylko wizualizacja
- Wyświetlanie par drzew – pierwotne oraz z proponowaną zmianą
- Użytkownik ocenia, czy zmiana jest prawidłowa
- System zapamiętuje te decyzje i nie pyta użytkownika po raz kolejny o tę samą zmianę
- „Efekt uboczny”: lista zmian, które należy wprowadzić w banku drzew

Bibliografia

- Cohn, T. i Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34, 637–674.
- Dickinson, M. i Meurers, W. D. (2003). Detecting inconsistencies in treebanks. W: *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, TLT '03, str. 45–56, Växjö, Szwecja.
- Dickinson, M. i Meurers, W. D. (2005). Prune diseased branches to get healthy trees! How to find erroneous local trees in a treebank and why it matters. W: *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, TLT '05, str. 41–52, Barcelona, Hiszpania.
- Kato, Y. i Matsubara, S. (2010). Correcting errors in a treebank based on synchronous tree substitution grammar. W: *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, str. 74–79, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Krasnowska, K., Kieraś, W., Woliński, M. i Przepiórkowski, A. (2012). Using tree transducers for detecting errors in a treebank of polish. W: *Proceedings of the 15th international conference on Text, Speech and Dialogue*, TSD '12, tom 7499 serii *Lecture Notes in Artificial Intelligence*, str. 119–126, Berlin. Springer-Verlag.
- Świdziński, M. i Woliński, M. (2010). Towards a bank of constituent parse trees for polish. W: *Proceedings of the 13th international conference on Text, Speech and Dialogue*, TSD '10, str. 197–204, Berlin, Heidelberg. Springer-Verlag.
- Wróblewska, A. (2012). Polish dependency bank. *Linguistic Issues in Language Technology*, 7(1).
- Wróblewska, A. i Przepiórkowski, A. (2012). Induction of dependency structures based on weighted projection. W: *Proceedings of the 4th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2012), Part I*, tom 7653 serii *Lecture Notes in Artificial Intelligence*, str. 364–374, Berlin. Springer-Verlag.