Automatyczne wykrywanie podmiotu zerowego

Mateusz Kopeć

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Seminarium ZIL, 27.01.2014

Plan prezentacji

Podmiot

Podmiot - obiekt w zdaniu w stronie czynnej, który:

- wykonuje czynność wyrażoną orzeczeniem (*Kot pije*),
- podlega procesowi wyrażonemu orzeczeniem (*Kot rośnie*),
- znajduje się w stanie wyrażonym orzeczeniem (*Kot leży*).

W zdaniu w stronie biernej natomiast obiekt, wobec którego:

- czynność jest wykonywana (*Kot został nakarmiony*).

Krócej:

- obiekt, którego czynność opisuje orzeczenie zdania,
- najważniejszy obiekt do zrozumienia zdania.

Uwaga

Zdanie = zdanie pojedyncze lub zdanie składowe

Podmiot zerowy (domyślny)

W językach:

- bałtosłowiańskich,
- romańskich (nie wszystkich),

można pomijać podmiot!

Przykład

(1) *Maria wróciła już z Francji. ØSpędziła tam miesiąc.*

Na to, czego brakuje, wskazują:

- forma orzeczenia (liczba pojedyncza, rodzaj żeński),
- kontekst (Maria),
- wiedza czytającego.

Uzupełnianie podmiotu zerowego

Wady występowania zjawiska podmiotu zerowego:

- utrudnienie właściwej analizy zdania,
- problem z tłumaczeniami maszynowymi,
- problem z analizą koreferencji.

Droga do eliminacji problemu:

1. automatyczne wykrycie, że w zdaniu występuje podmiot zerowy (wstawienie $\underline{\emptyset}$),
2. zbadanie formę orzeczenia i wstawić przed nim odpowiedni zaimek ($\underline{\emptyset} \rightarrow \textit{Ona}$),
3. analiza koreferencji do odkrycia powiązania z innymi obiektami ($\textit{Ona} \rightarrow \textit{Maria}$).

Częstość zjawiska

Dostępne dane:

- Polski Korpus Koreferencyjny [5] zawiera m.in. ręczną anotację podmiotów zerowych,
- orzeczenia bez podmiotu są oznaczone (*spędziła*).

Podmiot zerowy występuje dla:

- $\sim$ 30% orzeczeń w języku polskim (PKK),
- $\sim$ 30% orzeczeń w języku włoskim [10],
- $\sim$ 41% orzeczeń w języku hiszpańskim [10].

Podmioty zerowe są często koreferencyjne. Średnia liczba obiektów w łańcuchu koreferencyjnym w tekstach PKK:

- 3,56 ogólnie,
- 5,89 jeśli jeden z nich jest podmiotem zerowym.

Wykrywanie podmiotu a analiza składniowa

- Pełna analiza składniowa zdania powinna wykryć podmiot domyślny
- Pełen rozkład zdania dla niektórych języków jest trudny
- Skupienie się wyłącznie na podmiocie domyślnym powinno dać lepsze wyniki

Powiązane prace

Próby wykrywania podmiotu zerowego przy użyciu maszynowego uczenia się dla innych języków:

- brazylijski portugalski [9] – 77% czasowników ma podmiot, 21% używa podmiotu zerowego, 2% to konstrukcje bezosobowe. Cechy morfoskładniowe oraz informacje z parsera dają skuteczność 83,04%,
- hiszpański [8] – 71% czasowników ma podmiot, 26% używa podmiotu zerowego, 3% to konstrukcje bezosobowe. Skuteczność 87,6%,
- chiński [13] – tylko 3% podmiotów zerowych, ważenie przykładów 8:1, 50,9% F_1,
- rumuński [4] – 26,7% podmiotów zerowych, po zrównoważeniu liczności klas 74,5% skuteczności.

Problem

Uproszczona część mowy	Oryginalna część mowy	Tag	Liczba	Przypadek	Rodzaj	Osoba
Rzeczownik	Rzeczownik	subst	+	+	+	
	Forma deprecjatywna	depr	+	+	+	
	Liczebnik główny	num	+	+	+	
	Liczebnik zbiorowy	numcol	+	+	+	
	Odsłownik	ger	+	+	+	
	Zaimek nietrzecioosobowy	ppron12	+	+	+	+
	Zaimek trzecioosobowy	ppron3	+	+	+	+
Czasownik	Forma nieprzyszła	fin	+			+
	Forma przyszła *być*	bedzie	+			+
	Aglutynant *być*	aglt	+			+
	Pseudoimiesłów	praet	+		+	
	Czasownik typu *winien*	winien	+		+	

Tablica: Przyjęte uogólnienia części mowy na bazie NKJP [6]

Problem

Czy dany Czasownik ma podmiot w postaci Rzeczownika?

Odstępstwa od NKJP

Przyjęte uproszczenia w stosunku do NKJP:

- liczebniki, odsłowniki i zaimki są Rzeczownikami – mogą być podmiotami i posiadają informacje morfoskładniowe jak zwykłe rzeczowniki,
- nie traktuję *siebie* (tradycyjnie zaimek) jako Rzeczownika, ponieważ nie może być podmiotem,
- tagi: *impt, imps, inf, pcon, pant, pact, ppas, pred*, w NKJP czasownikowe, dla nas nie są Czasownikami, ponieważ nie mogą mieć podmiotu,
- nie zajmuję się wykrywaniem (rzadkich) podmiotów nierzeczownikowych, np.

  (2) *Niestety znaleźli się tacy*$_{adj}$.

Polski Korpus Koreferencyjny

Dane do badania:

- pochodzą z Polskiego Korpus Koreferencyjnego,
- obejmują 779 tekstów z 1773 w korpusie,
- zawierają ręcznie oznaczone (wśród innych wystąpień) czasowniki bez podmiotu,
- zawierają automatyczny podział na zdania, słowa i tagowanie morfoskładniowe (PANTERA [1]).

Przygotowanie danych

Problem

Czy dany Czasownik ma podmiot w postaci Rzeczownika?

Procedura przygotowania danych:

1. wyszukanie słów z tagami Czasowników w PKK (na podstawie tagów z PANTERY)
2. jeżeli takie słowo jest wystąpieniem, jest to orzeczenie bez podmiotu,
3. wpp. jest to orzeczenie z podmiotem.

Niedoskonałości w danych (automatyczne tagowanie):

- brak czasowników nie odnalezionych przez tager,
- nadmiarowe wystąpienia z błędnie przypisanym przez tager tagiem czasownikowym.

Rozmiar danych

Dane zostały podzielone na równoliczne zbiory:

- rozwojowy – do prac nad algorytmem,
- testowy – wykorzystany tylko raz, do końcowej ewaluacji.

Zbiór	# tekstów	# zdań	# tokenów	# czasowników	# wystąpień	# wystąpień czasownikowych
Rozwojowy	390	6481	110379	10801	37250	3104
Testowy	389	6737	110474	11000	37167	3106
Razem	779	13218	220853	21801	74417	6210

Tablica: Rozmiar danych

Zgodność anotatorów

- 210 tekstów w PKK posiada 2 niezależne ręczne anotacje
- Występuje tam 5879 Czasowników
- Zaobserwowana zgodność: 92,57%
- Zgodność przypadkowa: 57,52%
- κ Cohena: 82,51%

		Anotacja B	
		Podmiot zerowy	Podmiot zwykły
Anotacja A	Podmiot zerowy	1581	231
	Podmiot zwykły	206	3861

Tablica: Macierz koincydencji niezależnych anotatorów

Wyniki parsera zależnościowego

Przetestowałem parsowanie głębokie:

- Wykorzystałem parser zależnościowy Aliny Wróblewskiej [12], raportujący 71% LAS[1] i 75,2% UAS[2].
- Każdy Czasownik, z którego nie wychodziła relacja typu `subj`, został oznaczony jako nie posiadający podmiotu.
- Ta procedura pozwoliła osiągnąć:
  - Skuteczność: 67,23%,
  - Dokładność: 46,53%,
  - Kompletność: 90,47%
  - F_1: 61,45%.
- Nie jest to zadowalający wynik.

[1] Labeled attachment score – procent słów z poprawnym nadrzędnikiem i typem zależności

[2] Unlabeled attachment score – procent słów z poprawnym nadrzędnikiem.

Cel (na podstawie zbioru rozwojowego)

Plan minimum:

- Trywialny model (zawsze jest podmiot): 71,13%
- Parser zależnościowy: 67,23%

Plan maksimum:

- Zgodność anotatorów: 92,57%

Potencjalne trudności

Zadanie nie jest łatwe, bo:

1. język polski ma swobodny szyk,
2. zdarzają się podmioty niemianownikowe:

   (3) *Pieniędzy*$_{subst:gen}$ *nie starczy dla wszystkich.*
3. podział na zdania pojedyncze jest nietrywialny.

Zarys algorytmu

Maszynowe uczenie się polega na:

- Określeniu cech Czasownika i kontekstu, które mogą pomóc komputerowi w decyzji
- Dostarczenia wzorcowych danych, z opisaną decyzją dla konkretnych przykładów

Uczeniem można sterować poprzez karanie za błędy, niekoniecznie tak samo dla różnych rodzajów błędów:

- *false positive* – niesłusznie wykryty podmiot zerowy
- *false negative* – niewykryty podmiot zerowy

Podejście I - Algorytm wysokiej kompletności

- Minimalizacja błędów *false negative* – jak najmniej niewykrytych podmiotów zerowych
- Algorytm: oznaczać podmiot zerowy wszędzie tam, gdzie brakuje idealnego kandydata na podmiot w zdaniu pojedynczym zawierającym badany Czasownik
- Idealny kandydat na podmiot:
  - Rzeczownik w mianowniku
  - całkowita zgodność morfologiczna z Czasownikiem
- Podział zdania złożonego na zdania pojedyncze:
  1. podział po każdym: *i, albo, lub, „ ", ’,-,–,)*,
  2. łączenie sąsiadów aż każde zdanie pojedyncze będzie zawierało Czasownik.

Analiza błędów algorytmu wysokiej kompletności

- Kompletność: 97,7%
- Dokładność: 22,56%
- Liczba przykładów w zbiorze rozwojowym: 10801
- Liczba nie znalezionych podmiotów zerowych: 40

Kategoria błędu	Opis błędu	Liczność
Nie błąd	Pomyłka w anotacji ręcznej	8
Błąd tagera	Brak kropek na końcu zdania w tekstach mówionych	14
	Błędny przypadek (*nom* zamiast *acc*)	3
Błąd podziału na zdania poj.	Problem z podziałem narrator/mówca	2
	Zdanie poj. wydzielone tylko względem jednego *()* lub *””* z pary	2
	Brakujące przecinki w tekście	3
	Podział zdań na przecinku między przymiotnikami	1
Trudne przypadki	*jak/jako* Rzeczownik	5
	Gdy ocknął się drugi raz. . .	1
	Cały czas chodził . . .	1

Tablica: Analiza błędów FN algorytmu wysokiej kompletności

Wnioski

Poprawa podziału na zdania/zdania pojedyncze:

- Wymuszenie podziału na *?* i *!*,
- Zabronienie podziału wewnątrz grupy składniowej wykrytej przez płytki parser SPEJD [7],
- Wprowadzenie poszukiwania nie tylko wewnątrz zdania (które może być błędne), ale także w oknie wokół Czasownika.

Ciekawy problem:

- *jako/jak* i Rzeczownik w mianowniku:

(4) *. . . jako głównodowodzący armii Ø nie miał prawa nawet startować.*

Podejście II - Algorytm wysokiej dokładności

- Minimalizacja błędów *false positive* – jak najmniej niesłusznie wykrytych podmiotów zerowych
- Algorytm: maszynowe uczenie się, z 5 razy bardziej bolesną karą za błąd *fp* niż *fn*.
- Cechy brane pod uwagę przez klasyfikator:
  - tag poprzedniego/następnego słowa
  - obecność Rzeczownika zgodnego (w różnym stopniu) z Czasownikiem:
    - w zdaniu
    - w zdaniu pojedynczym
    - w oknie wokół Czasownika

Analiza błędów algorytmu wysokiej dokładności

- Kompletność: 26%
- Dokładność: 97,46%
- Liczba przykładów w zbiorze rozwojowym: 10801
- Liczba niesłusznie oznaczonych podmiotów zerowych: 21

Kategoria błędu	Opis błędu	Liczność
Nie błąd	Pomyłka w anotacji ręcznej	11
Błąd tagera	Błędny rodzaj/liczba/tag	3
	Nierozpoznany rzeczownik	1
Błąd podziału na zdania poj.	Zdanie rozpoczynające się od przecinka i zaimka *który/jaki*	2
Inne przypadki	Pseudo-czasowniki (nie wymagające podmiotu)	3
	Niedoskonałość modelu	1

Tablica: Analiza błędów FP algorytmu wysokiej dokładności

Wnioski

- Wyróżnienie pseudo-czasowników niewymagających podmiotu, np.

  (5) *Bywało, że niektórzy . . .*

- Przecinek przed *który/jaki* nie powinien dzielić zdania na zdania pojedyncze.

Podejście III - finalne

- Maksymalizacja skuteczności
- Algorytm regułowy RIPPER [2] z WEKI [3]
- Cechy opisane w dalszej części

Cechy Czasownika

- 3 cechy Czasownika:
  - czy jest na liście pseudo-czasowników (nie wymagających podmiotu) stworzonej na podstawie słownika składniowego czasowników polskich [11] (cecha binarna),
  - liczba Czasownika (cecha nominalna),
  - tag Czasownika (cecha nominalna).

Cechy dotyczące słów wokół Czasownika

- 6 cech słów wokół czasownika:
  - tag kolejnego słowa (cecha nominalna),
  - tag poprzedniego słowa (cecha nominalna),
  - czy poprzedni tag to *praet* – do pomocy w przypadkach:

    (6) $\ldots$ *była*$_{praet}$ *m*$_{aglt:pri}$ $\ldots$

    (cecha binarna),
  - czy wśród 2 poprzednich tagów jest pred – do przypadków:

    (7) *Można*$_{pred}$ *się było*$_{praet}$ *tego spodziewać.*

    (8) *Trzeba*$_{pred}$ *było*$_{praet}$ *myśleć wcześniej.*

    (cecha binarna),
  - czy następny tag to inf – do przypadków:

    (9) *Wtedy należy*$_{fin}$ *poprosić*$_{inf}$*.*

    (cecha binarna),
  - czy poprzednie słowo to przecinek (cecha binarna).

Cechy dotyczące długości zdania

- 2 cechy:
  - liczba słów w zdaniu (cecha liczbowa),
  - liczba słów w zdaniu pojedynczym (cecha liczbowa).

Cechy dotyczące obecności Rzeczownika

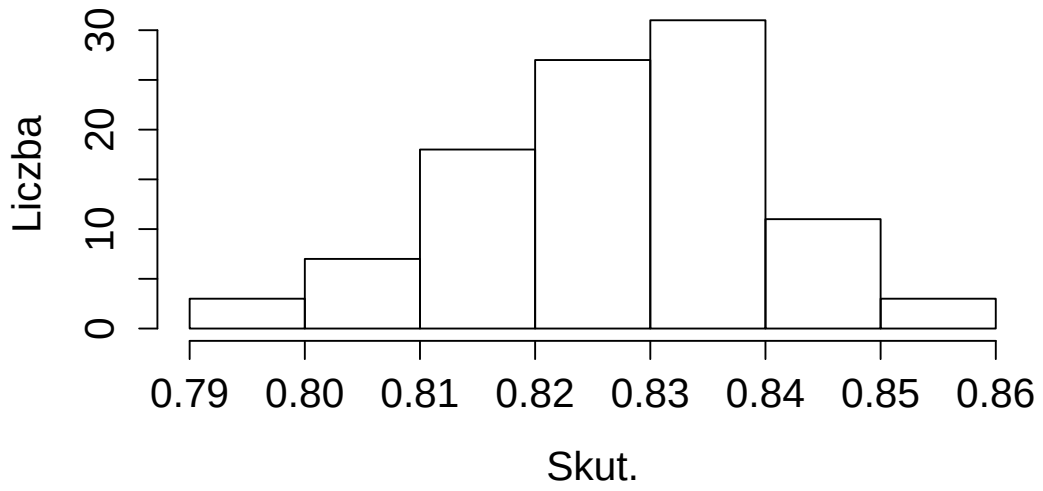
- 17*7 = 119 cech binarnych dotyczących obecności zgodnego Rzeczownika nie poprzedzonego *jak/jako* w oknie wokół Czasownika.
- Rozważane okna (17 możliwości):
  - zdanie pojedyncze,
  - całe zdanie,
  - okno od 1 do 5 słów przed Czasownikiem,
  - okno od 1 do 5 słów za Czasownikiem,
  - okno od 1 do 5 słów przed i za Czasownikiem.
- Rozważany stopnień zgodności (7 możliwości):
  - Rzeczownik w mianowniku – NOM,
  - Rzeczownik o zgodnej liczbie – NUM,
  - Rzeczownik o zgodnej osobie lub rodzaju – POG,
  - NUM i POG,
  - NUM i NOM,
  - POG i NOM,
  - NOM, NUM i POG.

Cechy dotyczące obecności 2 Rzeczowników

- 17*3 = 51 cech binarnych dotyczących obecności co najmniej 2 zgodnych Rzeczowników nie poprzedzonych *jak/jako* w oknie wokół Czasownika.
- Rozważane okna (17 możliwości):
  - zdanie pojedyncze,
  - całe zdanie,
  - okno od 1 do 5 słów przed Czasownikiem,
  - okno od 1 do 5 słów za Czasownikiem,
  - okno od 1 do 5 słów przed i za Czasownikiem.
- Rozważany stopnień zgodności obydwu (3 możliwości):
  - NOM,
  - POG,
  - NOM i POG.

Skuteczność na zbiorze rozwojowym

- 10-krotna walidacja krzyżowa
- 10 powtórzeń dla różnych seedów podziału przykładów na testowe/uczące
- średnia skuteczność: 82.74%, przedział ufności potwierdzony testem Shapiro: [82.49%, 82.99%].



Rysunek: Histogram skuteczności testowanej 100 razy

Porównanie wyników – zbiór rozwojowy

Algorytm	Skut.	Dokł.	Komplet.	F_1
Parser	67,2%	46,5%	90,5%	61,5%
Trywialny	71,1%	100,0%	0,0%	0,0%
Podejście I	44,6%	22,6%	97,7%	36,7%
Podejście II	78,4%	97,5%	26,0%	41,1%
Podejście III	**82,9%**	72,9%	64,4%	**68,4%**

Tablica: Porównanie wszystkich algorytmów na zbiorze rozwojowym

Ewaluacja na zbiorze testowym

Wyłącznie dalej opisane wyniki były badane na zbiorze testowym. Wyniki modelu RIPPER:

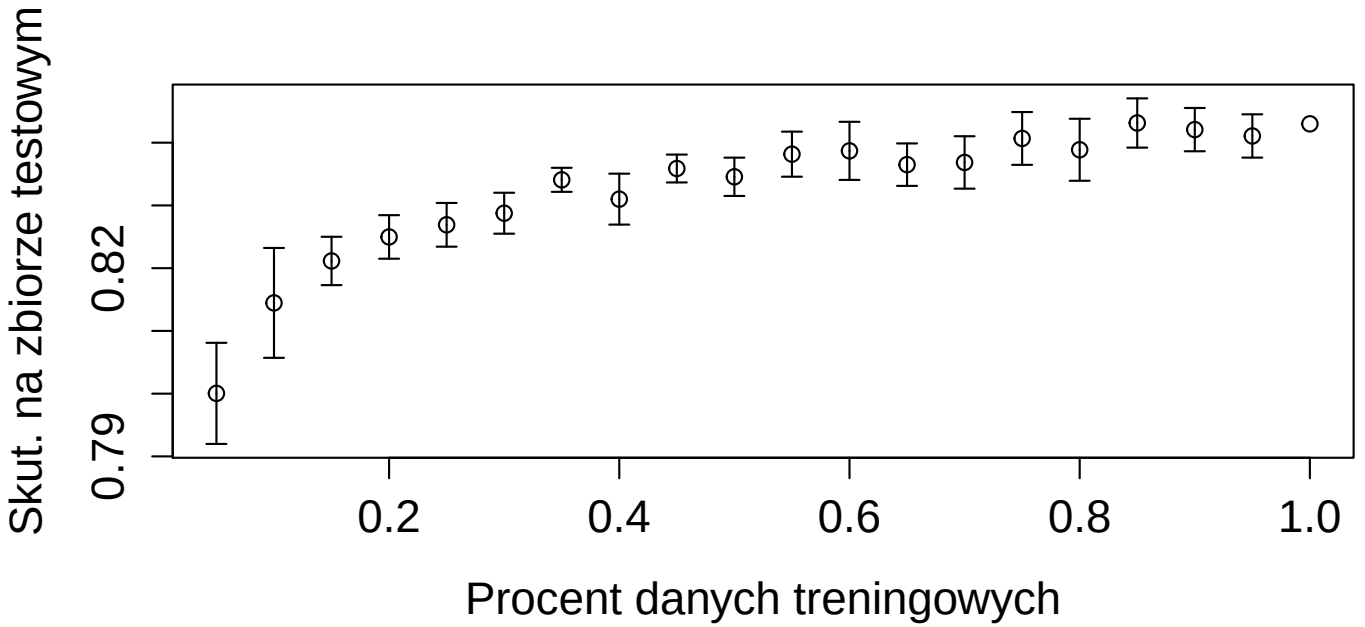
- Skuteczność : 83,38% (82,9% na testowym),
- Kompletność: 67,39% (64,4% na testowym),
- Dokładność: 71,97% (72,9% na testowym),
- F_1: 69,60% (68,4% na testowym).

		Prawda	
		Podmiot zerowy	Podmiot zwykły
Predykcje	Podmiot zerowy	2093	815
	Podmiot zwykły	1013	7079

Tablica: Macierz koincydencji modelu wytrenowanego na zbiorze rozwojowym i testowanego na testowym

Krzywa uczenia się

- Trenowanie: od 5% do 100% zbioru rozwojowego,
- Testowanie: zbiór testowy
- Każdy procent testowany 10 razy dla losowych przykładów
- Test Shapiro potwierdził rozkład normalny dla 16 z 19 prób



Rysunek: Krzywa uczenia się

Podsumowanie

Co zrobiłem:

- Zaprezentowałem automatyczny sposób wykrywania braku podmiotu przy użyciu maszynowego uczenia się.
- Wykorzystałem ważenie błędów do znalezienia trudnych przypadków i błędów w anotacji ręcznej.
- Zaprojektowałem heurystyczny podział na zdania pojedyncze.
- Skuteczność 83,38% istotnie przekroczyła 71,76%, jednak można zbliżyć się bardziej do 92,57%.

Przyszłość

Co można jeszcze zrobić:

- skupić się osobno na podziale na zdania pojedyncze,
- wykorzystać stworzone narzędzie jako etap wstępny w systemie analizy koreferencji,
- użyć techniki ważenia błędów do znalezienia błędów w ręcznej anotacji korpusu,
- przeanalizować powstałe reguły i wykorzystanie poszczególnych cech.

Koniec

Ø$_{sg:pri}$ Dziękuję za uwagę!

Podziękowania

Praca jest współfinansowana ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego. Projekt PO KL „Technologie Informacyjne: badania i ich interdyscyplinarne zastosowania” oraz projektu „Komputerowe metody identyfikacji nawiązań w tekstach polskich", finansowanego przez Narodowe Centrum Nauki (nr kontraktu 6505/B/T02/2011/40), 2011-2014.

Bibliografia I

Szymon Acedański.
A Morphosyntactic Brill Tagger for Inflectional Languages.
In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, Advances in Natural Language Processing, volume 6233 of Lecture Notes in Computer Science, pages 3–14. Springer, 2010.

William W. Cohen.
Fast effective rule induction.
In In Proceedings of the Twelfth International Conference on Machine Learning, pages 115–123. Morgan Kaufmann, 1995.

Bibliografia II

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten.
The WEKA data mining software: an update.
SIGKDD Explor. Newsl., 11(1):10–18, November 2009.

Claudiu Mihaila, Iustina Ilisei, and Diana Inkpen.
Zero Pronominal Anaphora Resolution for the Romanian Language.
Research Journal on Computer Science and Computer Engineering with Applications" POLIBITS, 42, 2011.

Bibliografia III

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska.
Polish Coreference Corpus.
In Zygmunt Vetulani, editor, Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 494–498, Poznań, Poland, 2013. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors.
Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish].
Wydawnictwo Naukowe PWN, Warsaw, 2012.

Bibliografia IV

Adam Przepiórkowski and Aleksander Buczyński.
Spejd: Shallow Parsing and Disambiguation Engine.
In Zygmunt Vetulani, editor, Proceedings of the 3rd Language & Technology Conference, pages 340–344, Poznań, Poland, 2007.

Luz Rello, Ricardo Baeza-Yates, and Ruslan Mitkov.
Elliphant: Improved Automatic Detection of Zero Subjects and Impersonal Constructions in Spanish.
In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 706–715, Avignon, France, April 2012. Association for Computational Linguistics.

Bibliografia V


Luz Rello, Gabriela Ferraro, and Iria Gayo.
A First Approach to the Automatic Detection of Zero Subjects and Impersonal Constructions in Portuguese.
Procesamiento del Lenguaje Natural, 49:163–170, 2012.


Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati.
Improving machine translation of null subjects in Italian and Spanish.
In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 81–89, Avignon, France, April 2012. Association for Computational Linguistics.


Marek Świdziński.
Syntactic dictionary of polish verbs, 1994.

Bibliografia VI


Alina Wróblewska.
Polish dependency bank.
Linguistic Issues in Language Technology, 7(1), 2012.


Shanheng Zhao and Hwee Tou Ng.
Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach.
In EMNLP-CoNLL, pages 541–550. ACL, 2007.