Opracowanie oprogramowania dla strukturalnego znakowania tekstów dygitalizowanych

Oresta Tymczyszyn
Politechnika Lwowska
Ukraina
oresta.tymchyshyn@gmail.com

Opracowanie tekstów

- edytowanie po rozpoznawaniu
- **strukturalne znakowanie**
- morfologiczno-syntaktyczna analiza
- semantyczna analiza

Automatyczny podział na zdania

- precyzyjne przeszukiwanie tekstów
- morfologiczno-syntaktyczna analiza
- dopasowanie korpusów równoległych
- automatyczne tłumaczenie

Termin ZDANIE

Zdanie w językoznawstwie termin ten oznacza wypowiedzenie służące do zakomunikowania jakiejś treści.

http://pl.wikipedia.org/wiki/Zdanie

Wypowiedzenie komunikat językowy wyrażony zespołem wyrazów powiązanych logicznie i gramatycznie lub jednym wyrazem

http://pl.wikipedia.org/wiki/Wypowiedzenie

Przyjmowane oznaczenie

Zdanie twór językowy odpowiadający samodzielnemu komunikatowi, granicami którego są:

- wielkie litery na początku
- kropka, wykrzyknik, znak zapytania, wielokropek, cudzysłów przed wielką literą, lub ich kombinacje na końcu

Niejednoznaczności lingwistyczne

Czy jest to jednym zdaniem ?

– Aha. Diabeł ulitował się nade mną – rzuciłem.

Stanisław Lem, Solaris

Niejednoznaczności lingwistyczne

A to ? ☺

– *A więc mam cię! Mam cię! Mój skarbie najdroższy, moje złoto, moje życie! Po tylu latach, po tylu trudach, po tylu niebezpieczeństwach – urywanym ze wzruszenia głosem mówił kapitan, wciąż jeszcze tuląc w objęciach to szlochającą to śmiejącą się żonę.*

Iwan Franko, Dla ogniska domowego

Niejednoznaczności lingwistyczne

«Він п'яний.

П'яний, як чіп», — подумав я і спалахнув гнівом.

Stanisław Lem, Solaris

Rwgułowa metoda

Michał Rudolf

Metody automatycznej analizy korpusu tekstów polskich

Warszawa: Uniwersytet Warszawski, Wydział Polonistyki, 2004

Program wychodzi z tego założenia, że akapit składa się ze zdań.

Otwartą ręką uderzyłem się lekko w twarz i powoli poszedłem do radiostacji. Gdy naciskałem klamkę, usłyszałem ostry głos:

- Kto tam?

Stanisław Lem, Solaris

Program wychodzi z tego założenia, że akapit składa się ze zdań.

Otwartą ręką uderzyłem się lekko w twarz i powoli poszedłem do radiostacji.

Gdy naciskałem klamkę, usłyszałem ostry głos:

- Kto tam?

Stanisław Lem, Solaris

Reguła 1

Od 1 stycznia do 30 sierpnia 1890 r.

, w ekspedycji uczestniczył również Feliks Koneczny, przebywający głównie w Rzymie, Florencji i w Wenecji.

Korpus IPI PAN

Reguła 1

Od 1 stycznia do 30 sierpnia 1890 r.

, w ekspedycji uczestniczył również Feliks Koneczny, przebywający głównie w Rzymie, Florencji i w Wenecji.

Korpus IPI PAN

Potencjalny znak końca, po którym następuje znak interpunkcyjny, nie jest znakiem końca

Reguła 1

Od 1 stycznia do 30 sierpnia 1890 r., w ekspedycji uczestniczył również Feliks Koneczny, przebywający głównie w Rzymie, Florencji i w Wenecji.

Korpus IPI PAN

Potencjalny znak końca, po którym następuje znak interpunkcyjny, nie jest znakiem końca

Reguła 2

– Gotów, Kelvin?

– rozległo się w słuchawkach.

Stanisław Lem, Solaris

Reguła 2

– *Gotów, Kelvin?*

– *rozległo się w słuchawkach.*

Stanisław Lem, Solaris

Potencjalny znak końca, po którym następuje mała litera (być może poprzedzona spacją lub myślnikiem), nie jest znakiem końca

Reguła 2

– Gotów, Kelvin? – rozległo się w słuchawkach.

Stanisław Lem, Solaris

Potencjalny znak końca, po którym następuje mała litera (być może poprzedzona spacją lub myślnikiem), nie jest znakiem końca

Reguła 3

O 17.

20: Jestem we mgle.

Stanisław Lem, Solaris

Reguła 3

O 17.

20: Jestem we mgle.

Stanisław Lem, Solaris

Kropka otoczona z obu stron cyframi nie jest znakiem końca

Reguła 3

O 17.20: Jestem we mgle.

Stanisław Lem, Solaris

Kropka otoczona z obu stron cyframi nie jest znakiem końca

Reguła 4

Niedźwiedź zastrzelony w Porąbce miał szerokość przedniej łapy (Sd) równą 12,5 cm i ważył 105 kg.

Korpus IPI PAN

Kropka poprzedzona skrótem pisanym bez kropki jest znakiem końca

Reguła 4

Problem: **omonimija skrótów**

- с. (село-wieś) – с (секунда-sekunda)
- m. (miasto) – m (metr)
- м. (місто-miasto) – м (метр- metr)

Reguła 4

Rozwiązanie: jeśli skrut jest poprzedzony cyframi – to jest skrót pisany bez kropki

Столицею України є ***м.*** *Київ.*

Конституція України

Легковий автомобіль протягом першої секунди руху пройшов шлях 0,25 м, протягом другої – 0,75 ***м.*** *З якою середньою […]*

І. Бакай, Збірник задач з фізики

Reguła 5

W ramach działalności międzykomitetowej Komisji Badań Współczesnych Ruchów Skorupy Ziemskiej, przy Wydziale III PAN, odbyły się posiedzenia, na których m.

in.

zostało przedstawionych i przedysku-towanuch kilka opracowań (prof.

Z. Kowalczyk, doc.

J. Niewiarowski, dr T. Wyrzykowski).

Reguła 5

[…], odbyły się posiedzenia, na których m.

in.

zostało przedstawionych i przedysku-towanuch kilka opracowań (prof.

Z. Kowalczyk, doc.

J. Niewiarowski, dr T. Wyrzykowski).

Kropka poprzedzona skrótem "nie końcowym" nie jest znakiem końca

Reguła 5

[…], odbyły się posiedzenia, na których m. in.

zostało przedstawionych i przedysku-towanuch kilka opracowań (prof. Z. Kowalczyk, doc. J. Niewiarowski, dr T. Wyrzykowski).

Kropka poprzedzona skrótem "nie końcowym" nie jest znakiem końca

Reguła 6

Po objaśnieniach J.

Potasz nawet laicy zaczęli się tłoczyć przy gablotach.

Korpus IPI PAN

Reguła 6

Po objaśnieniach J.

Potasz nawet laicy zaczęli się tłoczyć przy gablotach.

Korpus IPI PAN

Inicjał (wielka litera z kropką) nie kończy zdania

Reguła 6

Po objaśnieniach J. Potasz nawet laicy zaczęli się tłoczyć przy gablotach.

Korpus IPI PAN

Inicjał (wielka litera z kropką) nie kończy zdania

Wady metody regułowej

- konieczność wykorzystania wiedzy lingwisty
- nieuniwersalność (reguly są właściwe tylko dla badanego języka)

  lub z nie wielkimi zmianami dla grupy bliskich języków ☺

Zalety metody regułowej

- nie wymaga duzych ilosci danych treningowych
- nie wymaga informacji morfologicznej

Co daje możliwość zastosowania metody do języków z biednymi lingwistycznymi resursami

Produktywność

Program był sprawdzany ręcznie na tekście który składa się z 5545 zdań.

Wystąpiło 29 błędów, 16 z nich były spowodowane złym formatowaniem tekstu wejściowego. O takich błędach program sygnalizuje.