Narzędzia informatyczne w językoznawstwie

Wprowadzenie

Marcin Junczys-Dowmunt
junczys@amu.edu.pl

Zakład Logiki Stosowanej
http://www.logic.amu.edu.pl

10. października 2007

Informacje ogólne

Dyżury:

Poniedziałek, 11:45 - 13:15, CN 230

Kontakt:

E-Mail: junczys@amu.edu.pl

Tel.: +48 600 213 050

Materiały do wykładu:

Prezentacje do wykładów oraz polecenia do zadań domowych będą pojawiały się dzień po wykładzie na:

http://www.logic.amu.edu.pl/Narzędzia_informatyczne

Zasady zaliczenia

Wykład kończy się egzaminem

Pytania takie jak w zadaniach domowych

Wymagania minimalne

- Trzeba zebrać 60 punktów za zadania domowe.
- Zadania są wykonywane pisemnie.
- Zadania domowe trzeba oddać osobiście na wykładach.
- Nie można zdobyć więcej niż 10 punktów za jednym razem.

Jak ominąć egzamin?

- Zebranie 100 punktów za zadania domowe zwalnia z egzaminu z oceną "dobry".
- Każde dodatkowe 20 punktów podwyższa ocenę o pół stopnia.

Cele wykładu

- Będziemy się skupiać na językoznawstwie wspomaganym komputerowo (computer-aided linguistics), nie na językoznawstwie komputerowym (computational linguistics)
- Jednym z celów wykładu jest zaznajomienie słuchaczy technologiami, standardami oraz narzędziami informatycznymi stosowanymi w językoznawstwie
- Innym celem jest doprowadzanie słuchaczy do takiego stopnia zaawansowania, że będą w stanie sami stworzyć potrzebne narzędzia, jeżeli takie nie będą publicznie dostępne

Wstępny plan wykładu I

- Wprowadzenie do HTML i XHTML
- Praca na poziomie Wiersza poleceń
- Wprowadzenie do PERL
  - Podstawowe polecenia kontrolne
  - Podstawowe struktury danych
  - Operacje wejścia-wyjścia
- Przetwarzanie tekstów wielojęzycznych
  - Standardy kodowania i Unicode
  - Ujednolicanie kodowania
  - Kodowanie a Edytory tekstu
  - Kodowanie a (X)HTML (czyli kodowanie w internecie)
  - Kodowanie a PERL
- Wyrażenia regularne
  - Wyrażenia regularne w Edytorach tekstu
  - Wyrażenia regularne w PERL

Wstępny plan wykładu II

- XML
  - Opis a dane
  - Standardy opisu danych lingwistycznych
  - Edytory XML
  - Wyczytywanie, przetwarzanie, generowanie XML
  - DTD i XML-Scheme *
  - Transformacje XML czyli XSLT *
- Konsola Unix czyli Cygwin
- Korpusy (zasoby gotowe)
  - Korpusy dostępne w internecie
  - Narzędzia indeksowania i przeszukiwania korpusów
- Internet jako korpus
  - Wady i zalety internetu pojmowanego jako korpus
  - Wyszukiwanie danych lingwistycznych w internecie
  - Automatyczne "ściąganie" stron internetowych
  - Automatyczne czyszczenie stron

Wstępny plan wykładu III

- Niektóre metody kwantytatywne w lingwistyce korpusowej *
- Narzędzia do automatycznej annotacji lingwistycznej *
- Przechowywanie danych lingwistycznych *
- Elementy lingwistyki komputerowej *

Zadanie domowe

1.1. Ściągnąć i zainstalować w domu następujące programy:

EmEditor Free 6.00.3 Darmowa (i niestety okrojona) wersja edytora tekstu umożliwiającego wczytywanie dużych plików tekstowych. Podświetla składnie różnych języków, w tym HTML, XML, PERL, C++ itp.

www.emeditor.com/modules/download2/

ActivePerl 5.8.8.822 Dystrybucja języka skryptowego PERL dla wszystkich wersji Windowsa

www.activestate.com/Products/activeperl/

Część I

Dane lingwistyczne a komputer

Charakteryzacja danych lingwistycznych wg. G. Simonsa[1]

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- Wielojęzyczność danych lingwistycznych
- Sekwencyjność danych lingwistycznych
- Hierarchiczność danych lingwistycznych
- Wielowymiarowość danych lingwistycznych
- Wysoki stopień sprzężenia danych lingwistycznych
- Oddzielenie danych od formatu

[1] Simons, Gary F. 1998. The Nature of Linguistic Data and the Requirements of a Computing Enviroment for Linguistic Research. In *Using Computers in Linguistics: a practical guide*, John M. Lawler and Helen Arister Dry (eds.). London and New York: Routledge

Charakteryzacja danych lingwistycznych wg. G. Simonsa

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- **Wielojęzyczność danych lingwistycznych**
- Sekwencyjność danych lingwistycznych
- Hierarchiczność danych lingwistycznych
- Wielowymiarowość danych lingwistycznych
- Wysoki stopień sprzężenia danych lingwistycznych
- Oddzielenie danych od formatu

Wielojęzyczność danych lingwistycznych

- Każdy fragment tekstu wprowadzony do komputera jest wprowadzany w jakimś języku (naturalnym bądź formalnym)
- Dane z którymi pracujemy my językoznawcy zawierają często informacje w różnych językach

Przykłady

- Słowniki bilingwalne (układ równoległy)
- Rozprawy językoznawcze (układ zagnieżdżony)
- Korpusy równoległe
- Prace tanslatorskie
- Podręczniki
- ...

Przykłady - Słownik dwujęzyczny

Problemy wynikające z wielojęzyczności

- Problemy z wyświetlaniem tekstów wielojęzycznych
  - Problem brakujących lub niepoprawnych informacji o kodowaniu
  - Problem brakujących czcionek
  - Problem niedostosowania programu do wyświetlania tekstów wielojęzycznych
- Problemy z wprowadzaniem tekstów wielojęzycznych
  - Na klawiaturze jest tylko 105 klawiszy
  - Standardowa strona kodowa nie zawiera potrzebnych znaków
  - Program nie pozwala na zmianę ustawień kodowania
  - Program nie jest przystosowany do innych alfabetów, pism sylabicznych lub ideograficznych
- Problemy z przetwarzaniem tekstów wielojęzycznych
  - Komplikacje przy mieszaniu kodowań w jednym pliku
  - Wewnętrzna konwersja na kodowanie bardziej uniwersalne

Częściowe rozwiązania problemu

Jak było kiedyś ...

- Zestaw znaków ograniczony do 128 kodów (7 bitów)
- np. w kodowaniu ASCII: $65 \mapsto$ A , $66 \mapsto$ B , $126 \mapsto \sim$
- Różne rozszerzenia (8 bitówm, 256 kodów) np. ISO 8859-1, ISO 8859-2, CP 1250 ...
- Ponieważ było tylko 256 możliwych kodów, trzeba było zmieniać przyporządkowania

Nieco lepsze rozwiązanie problemu

Jak być powinno ... Unicode

- W tej chwili standard Unicode obejmuje 99 089 znaków
- Jest miejsce na ponad milion dalszych
- Numery są przydzielone grafemom, nie glyfom
- Istnieją plany włączenia wszystkich systemów znakowych
- Zawiera np. egipskie hieroglify, pismo Majów, pismo Rongorongo z Wysp Wielkanocnych (niezrozumiane) itp.
- Zawiera oprócz pism klasycznych inne systemy znakowe np. pismo Braille'a, alfabet IPA, symbole matematyczne itp.

Niektóre problemy pozostają

- W jaki sposób można wygodnie wprowadzić 99 089 znaków?
- Istnieje wiele tysięcy czcionek, ale mniej niż tuzin obejmuje większość standardu Unicode
- Trzeba nadal korzystać z wyspecjalizowanych czcionek, np. dla pisma chińskiego, pisma Rongorongo ...

Więcej o standardzie Unicode i historii kodowań na późniejszych wykładach

Charakteryzacja danych lingwistycznych wg. G. Simonsa

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- Wielojęzyczność danych lingwistycznych
- **Sekwencyjność danych lingwistycznych**
- Hierarchiczność danych lingwistycznych
- Wielowymiarowość danych lingwistycznych
- Wysoki stopień sprzężenia danych lingwistycznych
- Oddzielenie danych od formatu

Sekwencyjność danych lingwistycznych - w czasie

- Wypowiedzi i tekst są produkowane i odbierane w czasie
- Tzn. przy jakiejkolwiek segmentacji danych językowych na elementy (fonemy, litery, znaki, morfemy, wyrazy, frazy, zdania itd.) mamy określoną czasowe następstwo tych elementów
- Czyli mamy do czynienia z pewną sekwencje elementów (zależnych od wybranej segmentacji)

Sekwencyjność danych lingwistycznych - w przestrzeni

Gdy musimy przechowywać dane lingwistyczne, następstwo czasowe (konieczne dla człowieka) jest często reprezentowane w zupełnie inny sposób

Mowa nagrana analogowo

- Mowa w postaci danych analogowych na taśmie magnetycznej lub na płytach winylowych
- Dane lingwistyczne w postaci elektro-magnetycznej sekwencyjnie rozłożone na taśmie
- Dane lingwistyczne w postaci rowków na płycie winylowej, rozłożonych przestrzennie w formie spirali

Sekwencyjność danych lingwistycznych - w przestrzeni

Pismo tradycyjne

- Tekst pisany/drukowany na papierze: sekwencje przestrzenne z określonym kierunkiem, np. dla języków europejskich od lewej do prawej, od góry w dół, kartki są wertowane od przodu do tyłu
- Wiemy, że taki rozkład nie jest oczywisty np.

Japoński	książki wertowane od tyłu do przodu, tekst pisany od góry do dołu
Arabski	tekst pisany od prawej do lewej
...	

Sekwencyjność danych lingwistycznych - dane cyfrowe

- W ostateczności mamy jeden sposób segmentacji danych cyfrowych: na **bity i bajty**
- Również dane cyfrowe są sekwencyjne, jednak trudno mówić o kierunku zapisu (w zasadzie zależy od nośnika danych)
- Z dokładnością do kodowania (*poprzedni dział o wielojęzyczności*) możemy założyć, że tekst w każdym języku jest cyfrowo zapisywany w ten sam sposób, nieważne czy to Polski, Japoński czy Rongorongo
- Poprawne wyświetlanie tekstów w tych językach jest sprawą oprogramowania

Problemy wynikające z sekwencyjności

- Wydaję się oczywiste, że musimy zachować informacje o sekwencyjności na różnych poziomach (litery, wyrazy, zdania, akapity, rozdziały)
- Jednak reprezentacja sekwencyjna jest pewnie najmniej wydajna pod względem **wyszukiwania informacji**

Pytanie

- Czy indeks w książce jest zorganizowany według sekwencyjności treści tej książki?
- Odpowiedź: **NIE**
- Układ według innych kryteriów: tutaj alfabet i odsyłacze w postaci numerów stron

Problemy wynikające z sekwencyjności

Pytanie 2

- Wyobraźmy sobie pełny indeks, który zawiera wszystkie wyrazy książki i oprócz numerów stron, numery wierszy i pozycję wyrazu w wierszu. Czy taki indeks pozwala na odczytanie książki?
- Odpowiedź: **TAK**
- Ale będzie okropnie niewygodne i nieefektywne. Niemniej informacje o sekwencyjności zostały zachowane
- Za to potrafimy znaleźć każdy pojedyncze wyraz w dosyć szybki sposób (jednak nie jest to optymalny sposób)

Wyszukiwanie informacji

Można więc działać na dwa sposoby:

- Wyszukiwać sekwencyjnie
- Indeksować i szukać niesekwencyjnie

Wyszukiwanie sekwencyjne

- Polega na porównywaniu kolejnych elementów **tekstu** do **wzorca**
- Wyszukiwanie kończy się, gdy znajdziemy element pasujący do wzorca lub gdy dotrzemy do końca pliku
- Każde kolejne wyszukiwania zaczynamy od punktu wyjścia (niekoniecznie od początku tekstu)
- Ciekawe narzędzie: **wyrażenia regularne** w edytorach tekstu lub w PERLu

Wyszukiwanie według indeksu

- Tworzymy **jednorazowo** indeks do **wielokrotnego** użytku
- Polega na porównywaniu kolejnych elementów **indeksu** do wzorca (gdy mamy inteligentny indeks to możemy od razu znaleźć odpowiedni element)
- Wyszukiwanie kończy się, gdy znajdziemy element pasujący do wzorca (nastąpi skok do wyznaczonego miejsca w tekście) lub gdy indeks nie zawiera pasujących wpisów
- Kolejne wyszukiwania nie wymagają przemieszczania się do punktów wyjścia w tekście
- Narzędzia: darmowe programy indeksujące lub PERL

Która metoda jest lepsza?

To zależy od naszych potrzeb

- Gdy mamy wielkie zbiory tekstów lepiej jest indeksować
- Dla małych zbiorów nie zawsze się opłaca indeksować
- Za to wyrażenia regularne są o wiele bardziej elastyczne
- Za pomocą indeksów znajdziemy tylko te informacje, które zostały uwzględnione w trakcie budowy indeksu

Można też mieszać oba podejścia jak to ma miejsce w profesjonalnych narzędiach korpusowych, np. TigerSearch, Poliqarp

Charakteryzacja danych lingwistycznych wg. G. Simonsa

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- Wielojęzyczność danych lingwistycznych
- Sekwencyjność danych lingwistycznych
- **Hierarchiczność danych lingwistycznych**
- Wielowymiarowość danych lingwistycznych
- Wysoki stopień sprzężenia danych lingwistycznych
- Oddzielenie danych od formatu

Struktura danych lingwistycznych

- Dane lingwistyczne są mocno ustrukturalizowane
- Dotyczy to dane prymarne, które zbieramy i badamy
- Dotyczy to również dane opisowe zawierające nasze analizy i interpretacje.
- Jednym rodzajem struktury, to **hierarchia**
- Pozostałe to **wielowymiarowość** i **sprzężenie** danych lingwistycznych

Hierarchiczność danych lingwistycznych

Hierarchia jest jedną z podstawowych koncepcji w językoznawstwie.

Przykłady

- Analizy składniowe zdań (czyli drzewa składniowe)
- Analiza tekstu (np. podział na rodziały, które zawierają akapity, które zawierają zdania, które zawierają ...)
- Struktura słownika (wpisy składające się ze znaczeń, zawierające przykłady)
- ...

Hierarchia a sekwencyjność

- Niestety programy najbardziej rozpowszechnione (np. Microsoft Word) nie nadają się do przetwarzania opisów hierarchicznych
- Są one przystosowane do sekwencyjnego przetwarzania dokumentów
- Rodzaje hierarchii są ustalone z góry lub bardzo ograniczone (rodziały, podrozdziały, akapity i co dalej?)
- Dalsze poziomy hierarchii są zależne od formatu

Przykład - Słownik

abacus noun [L. abacus from Greek abax] *pl. -cuses, or -ci*

1. a frame with beads sliding back and forth on wires for doing arithmetics
2. in architecture, a slab forming the top of the capitol of a column

abaft adverb

1. in the direction of the stern, astern

abandon verb

1. to leave completely and finally; forsake utterly; desert: *to abandon one's farm; to abandon a child; to abandon a sinking ship.*
2. to give up; discontinue; withdraw from: *to abandon a research project; to abandon hopes for a stage career.*

Opis hierarchi za pomocą XML

- XML (ang. Extensible Markup Language) to uniwersalny język formalny przeznaczony do reprezentowania różnych danych w ustrukturalizowany sposób
- XML jest niezależny od platformy (Każdy dokument XML to tak naprawdę zwykły plik tekstowy)
- XML jest rekomendowany oraz specyfikowany przez organizację W3C.
- XML może być *przekształcany* do wielu innych formatów (np. HTML, DOC, PDF) za pomocą XSL (ang. Extensible Stylesheet Language)

Przykład - Słownik

abacus **noun** [L. abacus from Greek abax] *pl. -cuses, or -ci*

1. a frame with beads sliding back and forth on wires for doing arithmetics
2. in architecture, a slab forming the top of the capitol of a column

Przykład - Słownik XML

```
<entry>
    <headword>abacus</headword>
    <etymology>L. abacus from Greek abax</etymology>
    <paradigm>pl. -cuses, or -ci</paradigm>
    <sense n="1">
        <pos>n</pos>
        <def>a frame with beads sliding back and forth
             on wires for doing arithmetics</def>
    </sense>
    <sense n="2">
        <pos>n</pos>
        <def>in architecture, a slab forming the top
             of the capitol of a column</def>
    </sense>
</entry>
```

Przykład - Artykuł

Naukowcy alarmują: studia są groźne dla zdrowia

PAP 13:15

Wyprowadzenie się z domu i konieczność dostosowania się do nowych warunków życia i obcego otoczenia to wyzwania, przed którymi staje wielu młodych ludzi rozpoczynających studia. [...]

Artykuł na ten temat ukazał się w październikowym numerze "Journal of Youth and Adolescence".

Badania prowadzono wśród studentek University of Alberta. Część z nich - grupa pierwsza - pochodziła z odległych miejscowości, inne – grupa druga – przynajmniej pierwszy rok studiów spędziły w domu rodzinnym. [...]

Przykład - Artykuł XML

```
<document>
<title>Naukowcy alarmują: studia są groźne dla
zdrowia</title>
<agency>
    <name>PAP</name>
    <time>13:15</time>
</agency>
<summary>Wyprowadzenie się z domu i konieczność
dostosowania się do nowych warunków życia i obcego
otoczenia to wyzwania, przed którymi staje wielu
młodych ludzi rozpoczynających studia.</summary>
<paragraph n="1">Artykuł na ten temat ukazał się w
październikowym numerze "Journal of Youth and
Adolescence".</paragraph>
</document>
```

Specifikacje niektórych standardów za pomocą XML

XHMTL (ang. Extensible HyperText Markup Language) język służący do tworzenia stron WWW. XHTML jest następcą HTML

TEI (ang. Text Encoding Initiative) jest standardem elektronicznej reprezentacji tekstu wraz z informacją o jego treści

XCES (ang. XML Corpus Encoding Standard) wersja XML znanego standardu CES

TMX (ang. Translation Memory eXchange), standard elektroniczny służący do zapisu pamięci tłumaczeń

XML a przetwarzanie

- Niestety przetwarzanie XML nie jest zawsze łatwe
- Ale istnieją specjalne programy oraz moduły do PERL
- Jeden taki moduł nosi optymistyczną nazwę **XML::Simple**

Charakteryzacja danych lingwistycznych wg. G. Simonsa

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- Wielojęzyczność danych lingwistycznych
- Sekwencyjność danych lingwistycznych
- Hierarchiczność danych lingwistycznych
- **Wielowymiarowość danych lingwistycznych**
- Wysoki stopień sprzężenia danych lingwistycznych
- Oddzielenie danych od formatu

Wielowymiarowość danych lingwistycznych

- Zwykła sekwencja tekstu jest jednowymiarowa
- Dodając hierarchię pojawia się drugi wymiar
- Rożne sposoby interpretacji tekstu odpowiadają kolejnym wymiarom, np.
  - Intonacja w mowie
  - Znaczenie gramatyczne wyrazów
  - Semantyka
  - Pragmatyka
  - ...

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie
LEM:	Litwa	ojczyzna	mój	ty	być	jak	zdrowie

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie
LEM:	Litwa	ojczyzna	mój	ty	być	jak	zdrowie
POS:	PN	N	P	P	V	C	N

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie
LEM:	Litwa	ojczyzna	mój	ty	być	jak	zdrowie
POS:	PN	N	P	P	V	C	N
MORF:	voc:sg:f	voc:sg:f	voc:sg:f	nom:sg:f	3:pres:sg:f	-	nom:sg:n

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie
LEM:	Litwa	ojczyzna	mój	ty	być	jak	zdrowie
POS:	PN	N	P	P	V	C	N
MORF:	voc:sg:f	voc:sg:f	voc:sg:f	nom:sg:f	3:pres:sg:f	-	nom:sg:n
FON:	litvɔ	ɔjʧɨznɔ	mɔja	tɨ	jɛstɛɕ	jak	zdrɔvʲe

Przykład

SURF:	Litwo	Ojczyzno	moja	Ty	jesteś	jak	zdrowie
LEM:	Litwa	ojczyzna	mój	ty	być	jak	zdrowie
POS:	PN	N	P	P	V	C	N
MORF:	voc:sg:f	voc:sg:f	voc:sg:f	nom:sg:f	3:pres:sg:f	-	nom:sg:n
FON:	litvɔ	ɔjʧɨznɔ	mɔja	tɨ	jɛstɛɕ	jak	zdrɔvʲe
SEM:	COUNTRY ABSTRACTION		-	-	-	-	CONDITION

I znowu XML

```
<sentence>
    <word n="1">
        <surf>Litwo</surf>
        <lem>Litwa</lem>
        <pos>PN</pos>
        <morf>
            <case>vocative</case>
            <gender>femininum</gender>
            <number>singular</number>
        </morf>
        <fon>litvɔ</fon>
        <sem>country</sem>
    </word>
    <word n="2">
  ...
</sentence>
```

Charakteryzacja danych lingwistycznych wg. G. Simonsa

Każde narzędzie informatyczne działające na danych lingwistycznych musi brać pod uwagę następujące aspekty:

- Wielojęzyczność danych lingwistycznych
- Sekwencyjność danych lingwistycznych
- Hierarchiczność danych lingwistycznych
- Wielowymiarowość danych lingwistycznych
- **Wysoki stopień sprzężenia danych lingwistycznych**
- Oddzielenie danych od formatu