**Kraków, 16.05.2012**

**Modelowanie i symulacja komputerowa**

Projekt zaliczeniowy

Temat pracy:

***Oznaczanie części mowy***

Autor:

**Joanna Misztal**

1. **Cel projektu**

Stworzenie programu służącego do oznaczania części mowy z podanym zdaniu wykorzystując ukryte modele Markova

1. **Opis projektu**

**Wejście:**

Na wejściu do programu podane zostaje zdanie w postaci listy słów.

**Wyjście:**

Program zwraca najbardziej prawdopodobną sekwencję stanów (części mowy) dla słów z podanego zdania.

**Dane uczące:**

Zbiorem danych uczących jest zbiór zdań z poprawnie oznaczonymi częściami mowy.

**Zbiór stanów:**

Zbiorem stanów jest ponumerowana lista części mowy: rzeczownik, przymiotnik, czasownik, liczebnik, zaimek, przysłówek, przyimek, spójnik, wykrzyknik, partykuła

**Zbiór symboli wyjściowych(zaobserwowanych):**

Zbiór słów w słowniku (na podstawie danych uczących)

**Macierz prawdopodobieństwa:**

Macierz przejść między stanami zawiera prawdopodobieństwa dla bigramów, czyli prawdopodobieństwa występowania po sobie dwóch kolejnych stanów (części mowy) w zdaniu.

**Macierz prawdopodobieństwa emisji symboli:**

Prawdopodobieństwa emisji danego symbolu wyjściowego (słowa) dla danego stanu (części mowy).

**Wektor początkowy:**

Wektor prawdopodobieństwa wystąpienia danego stanu początkowego (danej części mowy jako pierwszego wyrazu w zdaniu).

**Opis działania:**

Na podstawie danych uczących tworzony jest opisany wyżej ukryty model Markova z podanymi parametrami.

Celem programu jest obliczenie najbardziej prawdopodobnej sekwencji stanów dla zaobserwowanych symboli wyjściowych, czyli podanie sekwencji części mowy dla słów w zdaniu.

Do tego celu użyty został dynamiczny algorytm Viterbiego szukający sekwencji stanów o maksymalnym prawdopodobieństwie wystąpienia.

1. **Wyniki:**

Nauczanie przeprowadzono na zbiorze 50 zdań w języku polskim, w których występowały wyrazy wieloznaczne (będące różnymi częściami mowy w zależności od znaczenia). Słownik składał się ze 163 słów. Zdania składały się w większości z 3-5 słów.

Testy przeprowadzono na 16 zdaniach złożonych ze słów ze słownika.

Błąd liczony jako stosunek liczba słów niepoprawnie oznaczonych do liczby wszystkich słów okazał się zależny od danych początkowych. Zdecydowanie lepsze wyniki zostały uzyskane, jeśli w początkowej macierzy prawdopodobieństw przyjmie się niezerowe prawdopodobieństwo dla wszystkich przejść.

Błąd waha się w okolicy 30%. Prawdopodobnie lepszy wynik zostałby uzyskany, gdyby użyty został model trigramów zamiast bigramów (prawdopodobieństwa dla 3 słów) i przy znacznie większym zbiorze danych uczących.

1. **Bibliografia:**

<http://pl.wikipedia.org/wiki/Cz%C4%99%C5%9B%C4%87_mowy>

<http://acl.ldc.upenn.edu/P/P99/P99-1023.pdf>

<http://www.aclweb.org/anthology-new/A/A92/A92-1018.pdf>

<http://people.cs.umass.edu/~mccallum/courses/inlp2004/lect10-tagginghmm1.pdf>

<http://en.wikipedia.org/wiki/Viterbi_algorithm#Pseudocode>