Przetwarzanie języka naturalnego
Ćwiczenia 3
Zajęcia 10

Zadanie 1. ($\star$) Przeplotem dwóch słów nazwiemy każde takie słowo, które powstało przez napisanie kilku znaków z pierwszego słowa, potem kilku z drugiego, znowu kilku z pierwszego i tak do wyczerpania obu słów. Przykładowo przeplotem słów *kotek* i *pies* może być *kopitesek*. Do gramatyki bezkontekstowej chcemy dodać nowy rodzaj produkcji $A \rightarrow_x BC$. Nieterminal A powinien generować wszystkie przeploty słów w_B oraz w_C, takich że w_B jest generowane przez B, a w_C jest generowane przez C.

Przedstaw bardziej formalną definicję języka generowanego przez tak wzbogaconą gramatykę. Czy klasa języków generowanych przez wzbogacone gramatyki bezkontekstowe jest równa klasie języków bezkontekstowych? Jaki ma to związek z językiem polskim?
UWAGA: gwiazdka nie mówi o trudności zadania, ale o tym, że do jego rozwiązania może potrzebna być wiedza niewymagana na naszym wykładzie, która nie będzie sprawdzana na egzaminie.

Zadanie 2. Rozważamy parser, który wyznacza nawiasowanie zdania. Czyli na przykład dla zdania:

Judyta wczoraj jechała szybkim pociągiem.

parser powinien zwrócić:

(Judyta wczoraj jechała (szybkim pociągiem))

Zaproponuj (i powiedz, jak je liczyć) dwie miary liczbowe jakości takiego rozbioru:

a) nie karzącą za zbyt szczegółowe nawiasowania (czyli *(Judyta ((wczoraj jechała) (szybkim pociągiem)))* też jest poprawne)

b) karzącą za wszystkie różnice.

Zadanie 3. Na wykładzie przy omawianiu algorytmu Bauma-Welcha zabrakło metody obliczania π_i (czyli prawdopodobieństwa rozpoczęcia w stanie i). Uzupełnij tę lukę.

Zadanie 4. Podczas realizacji algorytmu Forward-Backward dla dłuższych ciągów otrzymujemy małe liczby (zbyt małe dla typu float czy double). Zaproponuj jakiś sposób na poradzenie sobie z tym (inny niż przejście na liczby wymierne).

Zadanie 5. Jak wykorzystać model HMM w zadaniu korekty błędów ortograficznych? (opis powinien być dość dokładny, z dokładną interpretacją współczynników a oraz b, wraz z opisanym sposobem ich szacowania)

Zadanie 6. Na wykładzie przedstawiona była metoda obliczania wartości współczynników b_{ik} czyli indeksowanych stanem początkowym i symbolem emitowanym. Jak szacować te współczynniki w ukrytych łańcuchach Markowa, jeżeli chcemy, by zależały one dodatkowo od stanu końcowego.

Zadanie 7. Zdanie to ciąg słów. Zdanie zniekształcone to ciąg zbiorów słów (być może jednoelementowych). Funkcją zniekształcającą nazwiemy funkcję, która dla zdania $w_1 \dots w_n$ zwraca zdanie zniekształcone $W_1 \dots W_n$, takie że dla każdego i mamy $w_i \in W_i$. *Dezambiguacją* nazwiemy obliczanie przeciwobrazu funkcji zniekształcającej przeciętego ze zbiorem poprawnych zdań.

Załóżmy, że mamy daną gramatykę bezkontekstową, która generuje język polski. Jak ją wykorzystać do dezambiguacji? Co jeżeli nasza gramatyka opisuje tylko niektóre konstrukcje języka polskiego? Zastanów się nad praktyczną możliwością realizacji pomysłów z tego zadania.

Zadanie 8. Przypomnij, jak działa klasyfikator NBB. Jak można by go wykorzystać do następującego zadania: określić, czy w tekście dany rzeczownik rodzaju m3 (stół, dom, samochód) występuje w mianowniku (Stół stoi, samochód przyjechał), czy też w bierniku (widzę samochód, patrzę na dom, lubię mój stół).

Zadanie 9. Zaproponuj 6 reguł hipotetycznego tagera regułowego, które zajmują się następującymi dwoma słowami: *jak* oraz *miał*. Wszystkie Twoje reguły powinny odnosić się do mniej popularnego znaczenia tych słów. Postaraj się, by reguły odnosiły się do rzeczywistych sytuacji językowych, które znalazłeś w Internecie (powiedz, jak ich szukałeś, by każdy, kto chce to powtórzyć, miał łatwiej).

Zadanie 10. Powiedz, jak łączyć tager Brilla z innymi tagerami. Zaproponuj dwa scenariusze.

Zadanie 11. Dla tagera Brilla będziemy rozważać reguły postaci:

> jeżeli na pozycji -1 mamy t_1, na pozycji +1 mamy t_2 wówczas zmień na pozycji 0 t_3 na t_4

Teoretycznie przestrzeń możliwych takich reguł jest bardzo duża (liczba tagów do potęgi czwartej). Wyjaśnij, dlaczego w rzeczywistości takich reguł będzie dużo mniej.