Przetwarzanie języka naturalnego
Ćwiczenia 4
Pierwsze zajęcia w 2021 roku

Uwaga: Lista jest specjalnie krótsza, żeby zostało trochę czasu na nadrabianie zaległości z pracowni P3 i konsultację P4 (która do zajęć powinna mieć już finalną postać).

Przypominamy, że gwiazdka oznacza nieobowiązkowość materiału z zadania i niewliczanie zadania do maksimum. Nie ma związku z trudnością.

Zadanie 1. 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 2. Na wykładzie 12 (pierwsza część) mówiliśmy o tagowaniu wyrazów znacznikami IOB w celu wyznaczania płaskiego rozbioru. Przypomnij krótko ogólną ideę tego pomysłu. Powiedziane było również, że można tagować punkty między słowami. Wyjaśnij, jak zrealizować ten pomysł.

Zadanie 3. Zaproponuj jakieś podejście do zadania 1 z pracowni 4b.

Zadanie 4. Zaproponuj jakieś podejście do zadania 2 z pracowni 4b.

Zadanie 5. Napisz możliwie najprostszą gramatykę, która umożliwi parsowanie takich fraz jak: *krytyczna decyzja*, *silnik spalinowy*, *nowoczesny silnik spalinowy*, *wczorajsza awaria modułu napędowego autobusu elektrycznego*. Czy ta gramatyka jest jednoznaczna? (dlaczego?) Przedstaw wariant jednoznaczny tej gramatyki. Czy w ten sposób straciłeś możliwość „poprawnego semantycznie" rozbioru pewnych fraz?

Zadanie 6. Wśród fraz nominalnych występują w języku polskim połączenia typu *panem prezesem*, *panią dyrektor* albo *turkuciem podjadkiem*. Niemniej jednak nie jest uniwersalną regułą, że występowanie obok siebie dwóch rzeczowników o tym samym przypadku, liczbie i rodzaju jest zawsze frazą (na przykład: Mój **samochód stół** przewiózł, ale z szafą sobie nie poradził). Zaproponuj regułę, która (wykorzystując bigramy z korpusu i, być może, inne dane, pozwoli odróżnić pary rzeczowników, które na pewno nie tworzą takiej frazy od par rzeczowników, które taką frazę prawdopodobnie tworzą (przy założeniu występowania obok siebie).

Zadania na następną listę – nie deklarujemy ich teraz!

Zadanie 7. ⋆ Pokaż, że problem należenia słowa do języka generowanego przez gramatykę atrybutową (z atrybutami pochodzącymi ze skończonego zbioru) jest NP-trudny.

Wskazówka: (rot13.com): fcebohw mnxbqbjnp ceboyrz FNG. Flzobyr avrgrezvanyar tenzngIxv zbtn cemrpubjljnp vasbeznpwr b jnegbfpv cbqsbezhyl benm b jnegbfpvbjnavh jfmlfgxvpu mzvraalpu. Jlcebjnqmnal wrmlx zbmr olp qbfp geljvnyal.

Zadanie 8. (**2p**) Ograniczoną przez $p > 0$ gramatyką PCFG nazwiemy parę składającą z liczby p oraz gramatyki PCFG. Gramatyka (p, G) generuje te słowa, które są generowane przez G, dla których najbardziej prawdopodobne drzewo ma prawdopodobieństwo większe niż p.

Scharakteryzuj zbiór języków generowanych przez k-ograniczone PCFG. Jak widać (?) nie jest to szczególnie ciekawe pojęcie. Zaproponuj jakiś jego wariant, który wyda Ci się mniej trywialny. Uwaga: o gramatykach PCFG jest na wykładzie 12 (już opublikowanym)