Przetwarzanie języka naturalnego
Ćwiczenia 4
Zajęcia 10 i 14 stycznia

Uwaga: 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. Rozważmy następujący wariant gramatyki atrybutowej:

1. Produkcje wyglądają jak w GBK
2. Symbole nieterminalne mogą mieć co najwyżej 1 parametr.
3. Wartością tego parametru może być lista symboli terminalnych (kodowana tak jak listy w Prologu, czy Haskellu, czyli za pomocą konstruktora łączącego głowę listy i jej ogon, oraz specjalnej stałej oznaczającej listę pustą)

Pokaż, jak za pomocą takiego formalizmu rozpoznawać języki $\{ww|w \in Sigma^*\}$ oraz $\{ww^Rww^R|w \in Sigma^*\}$. Dla ułatwienia podajemy przykładowe produkcje takiej gramatyki:

```
S(Xs) -> A(a:Xs) B(b:Xs)
S(X:Xs) -> A(X:[]) B(Xs)
S(Xs) -> a S(a:a:Xs) b
```

Zadanie 3. Pokaż, że każdą gramatykę bezkontekstową da się przedstawić w postaci normalnej Greibach nie zmieniając języka akceptowanego przez tą gramatykę (przy założeniu, że słowo puste nie należy do tej języka generowanego przez tę gramatykę).

Zadanie 4. Połączenie reguł Bottom-Up Predict oraz Fundamental Rule (zob. wykład 11) nie daje kompletnego parsera. Jakiej reguły (reguł) brakuje? Zaproponuj sposób praktycznej realizacji takiego parsera.

Zadanie 5. Połączenie reguł Top-Down Predict oraz Fundamental Rule (zob. wykład 11) nie daje kompletnego parsera. Jakiej reguły (reguł) brakuje? Zaproponuj sposób praktycznej realizacji takiego parsera.

Zadanie 6. 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 7. 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).

Zadanie 8. $\star$ Co to są algorytmy ewolucyjne? W jaki sposób mogą być pomocne przy generowaniu poezji, tak jak w zadaniu z Panem Tadeuszem?

Zadanie 9. ⋆ 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 tenznglxv zbtn cemrpubjljnp vasbeznpwr b jnegbfpv cbqsbezhyl benm b jnegbfpvbjnavh jfmlfgxvpu mzvraalpu. Jlcebjnqmnal wrmlx zbmr olp qbfp geljvnyal.

Zadanie 10. (**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 będzie na pierwszym wykładzie w 2019 roku.