Indukcja gramatyki z generowaniem przykładów

ewolucja i aktywne uczenie się

autor Łukasz Stafiniak, Marcin Grabowski

Etap pierwszy – generowanie populacji początkowej

Gramatyki w postaci Chomsky'ego, podejście „kompresja":

- nowe nieterminale generowane z częstych bigramów
- reguły rekurencyjne generowane, gdy rozkład z trigramów po podstawieniu nieterminala jest podobny do oryginalnego rozkładu bigramowego nieterminala: $A \rightarrow A\ B$ gdy częstość trigramów postaci $C\ A\ B$ i $A\ B\ D$ jest podobna do częstości odpowiadających bigramów $C\,A$ i $A\,D$
- przechowujemy populację gramatyk, osobno kompresujemy korpus każdą gramatyką
- spośród najczęstszych bigramów wybieramy te najrzadziej wybierane (w innych gramatykach)

W przyszłości warto też opracować wersję dla „dependency grammars".

Etap drugi – schemat algorytmu ewolucyjnego

Pomysł polega na konstrukcji korpusu zawierającego pozytywne i negatywne przykłady.

- Z każdej gramatyki generujemy przykładowe zdanie / próbkę zdań.
- Użytkownik ocenia poprawność składniową zdań z mini-korpusu wygenerowanego przez gramatyki z populacji.
- Daną gramatykę oceniamy licząc np.

$$\text{fitness} = \alpha p + \beta n$$

gdzie p to **true positives**, n to **true negatives**, $\alpha < \beta$, „negatives" to zdania których nie udało się sparsować tą gramatyką.

- Prowadzimy ewolucję do ustabilizowania się dostosowania, następnie generujemy kolejny mini-korpus, który dołączamy do zbioru uczącego (dyskontując dotychczasowy korpus, żeby zwiększyć „ruchliwość" ewolucji).

Operatory genetyczne

- elementy podejścia „Pittsburgh”: osobnik jest zbiorem reguł, krzyżowanie wymienia reguły
- operatory ukierunkowanej mutacji:
  - **usuwanie reguły**: usuń jedną z reguł użytych w derywacji jakiegoś „false positive”
  - **kreacja reguły**: dodaj regułę domykającą / poszerzającą najszerszy łuk w derywacji jakiegoś „false negative”
- elementy podejścia „Michigan”: można oceniać poszczególne reguły w oparciu o derywacje w których uczestniczą

Etap trzeci – gramatyki probabilistyczne

- Przed wprowadzeniem prawdopodobieństw, generowanie zdań do mini-korpusu ma charakter prowizoryczny.
- Gramatyki probabilistyczne pozwalają wybrać najbardziej prawdopodobne zdanie / zdania wypróbkowane z danej gramatyki.
- Być może warto generować gramatyki w formacie systemu **Alchemy**.
    - Wtedy jednym z celów projektu byłoby zapoznanie się z tym systemem.
- Indukcja gramatyki w oparciu o słowa, nie tylko tagi.
    - indukcja kategorii semantycznych