Parsing zależnościowy i nie tylko

Paweł Rychlikowski

Instytut Informatyki UWr

7 lutego 2019

Precision/Recall w ocenie parserów (przypomnienie)

- Precyzja:

$$P = \frac{\text{Liczba poprawnie odgadniętych fraz w zwróconym rozbiorze}}{\text{Liczba fraz w zwróconym rozbiorze}}$$

- Kompletność:

$$R = \frac{\text{Liczba poprawnie odgadniętych fraz w zwróconym rozbiorze}}{\text{Liczba fraz we wzorcowym rozbiorze}}$$

F_1-score

Gdy chcemy scharakteryzować rozbiór jedną liczbą, używamy średniej harnonicznej P i R,

$$F_1 = (\frac{\frac{1}{P} + \frac{1}{R}}{2})^{-1} = \frac{2PR}{P+R}$$

Inne sposoby oceny

Sposób 1

Czasem rozważa się dwa warianty oceny – z uwzględnieniem nazw fraz oraz bez uzwględniania.

Sposób 2

Cross-brackets – liczba fraz w których wzorcowy parsing i niewzorcowy mają niezgodne nawiasowania.

Zadania z parsowaniem NP

- Prostsza metodologia oceny:
  1. Mamy dobre frazy (frazy NP ze Składnicy)
  2. Mamy sztucznie wygenerowane złe frazy, przejrzane przez studentów.
- Trzeba odróżniać jedne od drugich

[obejrzyjmy plik np_original.pl]

State of the art w parsingu

- Najlepsze parsery frazowe dla języka angielskiego osiągają obecnie ponad 93% (F_1-score, z uwzględnieniem nazw nieterminali)
  - *Grammar as a Foreign Language*, Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, Geoffrey Hinton
- Dla języka polskiego – to trudniejsze pytanie. Istnieje praca (Woliński, Rogozińska), która podaje 94.1% skuteczności.
- Wynik niestety nie jest do końca miarodajny, bowiem korzysta m.in. ze Świgry.

Uwaga

Obecnie w zastosowaniach przemysłowych króluje raczej parsing zależnościowy (o którym dzisiaj będzie)

Czym jest PCFG dla gramatyki prawostronnej?

Definicja

Gramatyka bezkontekstowa jest prawostronnie liniowa, jeżeli jej produkcje mają postać: $A \to w_1 \dots w_n B$ oraz $A \to w_1 \dots w_n$

- **Fakt:** gramatyki prawostronnie liniowe definiują klasę języków regularnych.
- Lewostronna PCFG w postaci Chomskiego jest (w zasadzie) ukrytym łańcuchem Markowa, w którym nieterminalom odpowiadają stany ukrtyte/tagi.

Produkcja $T \to^p wT'$ mówi o prawdopodobieństwie $P(w, T'|T)$. Prawdopodobieństwa $P(w|T)$ i $P(T'|T)$ otrzymujemy sumując po wszystkich stanach lub po wszystkich słowach.

Generacja wykorzystująca PCFG

Uwaga

Przypominamy: PCFG $\approx$ treebank

- W wersji liniowej wyobrażamy sobie model bigramowy dla tagów i generowanie wypełnienia treścią po wylosowaniu tagów.
- W wersji standard PCFG możemy najpierw generować drzewko tagów, a następnie wypełniać je słowami.
- Czym jest symbol nieterminalny (przypominamy: $P(A \rightarrow \gamma|A)$):
  a) Prosty symbol nieterminalny: **fwe**
  b) j.w. + podstawowa informacja gramatyczna: **fwe:mos:poj:3**
  c) j.w. + informacja o typie: **fwe:mos:poj:3:[advp,np[bier],subj]**
  d) j.w. + informacja o nagłówku: **fwe:mos:poj:3:[advp,np[bier],subj]|nosił**

Generacja wykorzystująca PCFG. Problemy

- Powtarzające się reguły:
  przez spłoszonego niedoświadczonego spłoszonego niedoświadczonego terrorystę płacili śmy składki na nagrody na to samo najważniejsze pytanie na siemiradzkiego na komisariat
- Typy wyrazów:
  - myśli, że jedzie do krakowa vs
  - myśli, że jedzie do wałbrzycha vs
  - jedzie, że myśli do krakowa
  - Słowa krakowa oraz wałbrzycha są na pozycjach *podstawialnych*.
- Zbyt „fantazyjne" kształty drzewek – a my lubimy proste zdania, bo takie nam (ludziom) łatwo sparsować.

Dedukcja PCFG bez treebanku

Czy w ogóle jest możliwa?

- HMM to jakby dedukcja lewostronnej PCFG (przy określonych założeniach)
- Da się to uogólnić na gramatykę w postaci Chomskiego (algorytm Inside-Outside), opisany w podręczniku Statistical Natural Language Processing (Manning, Schuetze)... *ale o nim nie będziemy mówić*

Rozwiązanie heurystyczne

- Szukamy możliwie małej gramatyki, generującej wszystkie zdania.
  - Może zawierać fragmenty rzeczywistych gramatyk, które znamy lub umiemy łatwo napisać.
- Dla każdego zdania zgadujemy parsing (czyli losujemy, premiując jakoś parsingi wykorzystujące *lepsze* reguły)
- Następnie powtarzamy fazy EM:
  1. E: szacujemy prawdopodobieństwa produkcji
  2. M: znajdujemy optymalny parsing dla każdego zdania

Rozwiązanie heurystyczne (2)

- Jak mamy więcej parserów „starszej generacji” to możemy nimi parsować duży korpus i dodawać drzewo do treebanku, jeżeli parsery zgadzają się co do rozbioru.
- Takie podejście było stosowane między innymi w pracy Grammar as a Foreign Language (czyli nie dotyczy ono tylko PCFG).

Szczypta poezji

Zadanie z Panem Tadeuszem, w którym mocno korzystamy z prawdziwych zdań.

- Mimo różnych możliwości generowania tekstu w języku naturalnym ciągle istotne jest generowanie na podstawie prawdziwych zdań (w praktyce nic innego nie działa w niebazującej na regułach rozmowie na tematy ogólne).
  - (w tłumaczeniu też mamy prawdziwy tekst po jednej stronie)
- Wektory dla form bazowych są użyteczne, bo formy bazowe są częstsze niż słowa, czyli mamy więcej kontekstów dla jednego wektora.
- Stosujemy prostą sztuczkę, pozwalającą nam na uniknięcie problemów z wielobazowością (zobaczmy jaką).
- i przy okazji rzućmy okiem na plik rytmiczne-zdania-z-korpusu.txt

Zadanie poetyckie

Jak upoetycznić zdanie:
rozpoczęto budowę nowego zaplecza [*] i szatni na stadionie za kinem syrena . wynik:
rozpoczęto budowę nowego +okienka [*] i szatni na stadionie za kinem +syrenka .

Deppendency parsing

- Nie myślimy o frazach, lecz o relacjach między słowami w zdaniu (oznaczane strzałkami).
- Rozbiór to graf skierowany, którego węzłami są słowa, który spełnia pewne warunki
  1. Jest drzewem (acykliczny, spójny)
  2. Strzałki mówią:

     Ja jestem ważniejszy, a ty mnie opisujesz.

Deppendency parsing (2)

Przykładowo, dla zdania Jasiu mieszkał w pięknym domu mamy strzałki:

mieszkał $\rightarrow$ *Jasiu*
mieszkał $\rightarrow$ *w*
w $\rightarrow$ *domu*
domu $\rightarrow$ *pięknym*

Drzewo rozbioru

- Oczywiście możemy taki graf narysować jako drzewo (najczęściej dodajemy sztuczny węzeł **root**)
- Możemy też go zapisać jako term:

  mieszkał(jasiu, w(domu(pięknym)))

- W grafie możemy nazywać krawędzie: podmiot, dopełnienie, przydawka, ...

Uwaga notacyjna

Gdy piszemy kobieta(piękna) nie mówimy, o pozycji głowy wobec dziecka (dzieci). A ona czasami jest istotna (dlaczego?). Możemy zatem się umówić na notację z gwiazdką i pisać (czasami) kobieta(*,piękna), kobieta(piękna,*), i(jaś,*,małgosia), etc.

Nieco trudniejsze przypadki

Narysujmy rozbiory następujących zdań

1. Jaś i Małgosia nie lubią jeść pierników
2. Powiedział: I ty jesteś przeciwko mnie, Brutusie i umarł.
3. Niemniej jednak w pracy wszyscy mnie chwalą.
4. Słoń, którego trąba jest wilgotna, nie lubi truskawek.

Projekcyjność

- Intuicyjnie projekcyjność mówi:

  Każde poddrzewo grafu odpowiada spójnej frazie

- Istnieje wiele naturalnych rozbiorów, które nie są projekcyjne:

  wpłynąłem na suchego przestwór oceanu ::
  wpłynąłem(na(przestwór(oceanu(suchego))))

Projekcyjność

- Intuicyjnie projekcyjność mówi:

  Każde poddrzewo grafu odpowiada spójnej frazie

- Istnieje wiele naturalnych rozbiorów, które nie są projekcyjne:

  wpłynąłem na suchego przestwór oceanu ::
  wpłynąłem(na(przestwór(oceanu(suchego))))

Projekcyjność (2)

- Wydaje się, że projekcyjność jest „domyślną opcją” w wielu językach.
- Niemniej, zdania nieprojekcyjne się zdarzają:

Język	NPD	NPS
Niderlandzki	5.4%	36.4%
Niemiecki	2.3%	27.8%
Czeski	1.9%	23.2%
Słoweński	1.9%	22.2%
Portugalski	1.3%	18.9%
Duński	1.0%	15.6%

NPD – non projective deppendencies, NPS – non projective sentences

Projekcyjność

- Intuicyjnie projekcyjność mówi:

  Każde poddrzewo grafu odpowiada spójnej frazie

- Istnieje wiele naturalnych rozbiorów, które nie są projekcyjne:

  wpłynąłem na suchego przestwór oceanu ::
  wpłynąłem(na(przestwór(oceanu(suchego))))

Definicja

Krawędź $w_i \rightarrow w_j$ jest projekcyjna, jeżeli dla każdego słowa w_k pomiędzy w_i a w_j w zdaniu istnije ścieżka $w_i \rightarrow^* w_k$.

Problem z wieloznacznością słów

- Słowa (jak wiemy) są wieloznaczne, co powoduje pewne problemy. Przykłady:

  W stylu dość pięknym tonie ten statek.

- Każde z połączeń: tonie(pięknym) oraz tonie(statek) z osobna wygląda dobrze...
- ale razem nie mogą pojawić się w jednym rozbiorze!'

Problem z wieloznacznością słów (2)

- Inny przykład, z pozornie jednoznacznym słowem w (zastanówmy się, gdzie tkwi problem?)

  Stefan w domu przeczytał książkę.

- Chodzi o dobór tagów: zarówno w domu (loc) jak i w książkę acc) to sensowne połączenia (vide: Krytyk uderzył w książkę ostrzem swojej recenzji.)
- Podobnie: trzeba zdecydować się na jeden z tych wariantów i drugiego zabronić.

Problem z wieloznacznością słów. Próba rozwiązania

Możliwe są dwie opcje:

1. Przed rozbiorem ujednoznaczniamy tekst (na przykład za pomocą algorytmu tagowania)
2. Myślimy o tym, że rozbiór to jednocześnie utworzenie grafu i wybór parametrów dla węzłów (parametrami mogą być na przykład tagi).

Współcześnie wygrywa **opcja druga**.

Universal Deppendenciec

- Dane o rozbiorach (i POS-tagach) w ujednoliconym formacie, dla olbrzymiej liczby języków.
- Prosty, czytelny format (CONLL) – przykład za chwilę
- Popatrzmy na język korpusy z języka polskiego (sprawdź: że mają szlaban w korpusie LFG)

Czy umiemy sparsować:

Radikálnou inováciou je hláskoslovie a tvaroslovie.

Uwaga

Są możliwości łączenia korpusów – możemy czegoś się dowiedzieć o parsingu po polsku patrząc na słowacki.

Deppendency parsing. Algorytm dynamiczny, wersja projekcyjna

- W przypadku parsingu projekcyjnego też tworzą się frazy, odpowiadające poddrzewom rozbioru.
- To są specyficzne poddrzewa, w których możemy pominąć pewne „zewnętrzne" gałęzie na najwyższym poziomie. Przykładowo dla

  *miała(babuleńka, *, koziołki(dwa,rogate,*))*

  poddrzewami są

  *miała(babuleńka, *)*
  miała(, koziołki(dwa,rogate,*))*
  koziołki(rogate,)*

- Parsing polega na łączeniu takich fraz (poddrzew) ze sobą.

Algorytm dynamiczny, wersja projekcyjna

- Odpowiedni struktury U (dla CYK) pamięta teraz dla każdego przedziału informację o najlepszych drzewach zakorzenionych we wszystkich słowach.
- Jak łączymy dwie takie struktury, to musimy rozważyć połączenie każdego słowa z lewej z każdym z prawej (w obie strony) i wybrać takie, które maksymalizuje

$$v(w_1) + v(w_2) + \max\{v(w_1(*, w_2), v(w_2(w_1, *), \}$$

gdzie w_1 jest z lewej części, a w_2 – z prawej.

Złożoność

Niestety złożoność jesst dość duża: (jaka?) algorytm pracuje w czasie $O(N^5)$.
Da się to poprawić: Algorytmem Elsnera (być może na ćwiczeniach).

Algorytm Chu-Liu Edmonds

CLU działa wg następującego schematu:

1. Każdy wybiera ulubionego tatusia
2. Jak nie ma cyklu – ok.
3. Jak jest cykl, to zamieniamy cykl na jeden sztuczny węzeł, rekurencyjnie rozwiązujemy mniejsze zadanie, potem rozbijamy cykl.

Złożoność

Algorytm działa w czasie $O(N^2)$

Projekcyjne vs nieprojekcyjne

- Szukanie drzew nieprojekcyjnych jest szybsze.
- Wiemy, że algorytm projekcyjny w wielu sytyacjach nie znajdzie poprawnego rozbioru.
- Z drugiej strony wydaje się, że warto preferować rozbiory projekcyjne

Uwaga

Częściowo rozwiązuje ten problem preferowanie połączeń niezbyt odległych.

Problemy

- Problem z kontrolą liczby strzałek (zdanie ma 1 podmiot)
- Problem z 'Jasiem i Małgosią' (jak wyglądać ma rozbiór zdania Jaś i Małgosia zjedli pierniki). Opcje:
  1. zjedli(i(jaś,małgosia), pierniki)
  2. zjedli(jaś, małgosia, pierniki)
  3. zjedli(jaś(i(małgosia)), pierniki)
- W pierwszym przypadku mamy niezbyt ciekawe połączenie zjedli(i), w drugim – pomijamy i, a ponadto mamy połączenia: zjedli(jaś) oraz zjedli(małgosia), w trzecin pozostaje problen niezgodności liczby, do tego Jaś i Małgosia nie są równo traktowani.

Deterministyczny parsing

- Motywacja: nasz wewnętrzny parser (raczej) nie wykonuje żadnego dynamicznego algorytmu $O(N^5)$, a mimo to rozumiemy, co do nas mówią (na ogół)
- Wydaje się, że o rozbiorze podejmujemu dezycję lokalnie, czasem – raczej rzadko – wykonując coś w rodzaju płytkiego nawrotu.

Uwaga

Istnieje dużo algorytmów tego typu, przeglądających wejście od lewej do prawej i tworzących przyrastająco rozbiór. Oczywiście są one szybkie i (co ciekawe) wcale niekoniecznie gorsze od wolniejszych optymalizacyjnych.

Garden path sentences

Można „zhakować” nasz wewnętrzny parser za pomocą specjalnych zdań (tzw. garden patsh sentences)

The old man the boat.

Rozbiorem jest:

man(old(the),boat(the))

ale problem z rozbiorem jest taki, że bardzo nas sugeruje pojawienie się ciągu wyrazów the old man