Podsumowanie HMM. Gramatyki

Paweł Rychlikowski

Instytut Informatyki UWr

11 grudnia 2018

Algorytm Viterbiego

Dla danego ciągu obserwacji i modelu zwraca optymalną sekwencję stanów (czyli taką, która ma największe prawdopodobieństwo.

- Mówiliśmy w kontekście tagowania i rekonstrukcji samogłoskowej.
- Używa składowej bigramowej ($\log p(t_{i+1}|t_i)$) oraz składowej unigramowej ($\log p(w_i|t_i)$)
- Dynamicznie maksymalizuje sumaryczny zysk.

3 algorytmy HMM (2)

Algorytm Forward-Backward

Dla danego ciągu obserwacji i modelu zwraca rozkład prawdopodobieństwa stanów ukrytych dla każdego momentu historii

- Można wybrać najlepszy stan w każdym momencie. Wówczas mamy zwróconą sekwencję najlepszych stanów
- Obliczamy następujące prawdopodobieństwa (pomijamy *pod warunkiem* μ):

$$\alpha_i(t) = P(o_1 \dots o_{t-1}, X_t = i)$$

oraz

$$\beta_i(t) = P(o_t \dots o_T | X_t = i)$$

- α liczymy **w przód**, natomiast β – **w tył**

3 algorytmy HMM (2)

Algorytm Bauma-Welcha

Dla danego ciągu obserwacji i rodziny modeli zwraca model HMM z tej rodziny najlepiej tłumaczący obserwację

- Korzysta z algorytmu Forward-Backward
- Jest algorytmem EM, czyli przeplata dwie fazy:
  1. **M**: algorytm FB + dodatki (próba wyjaśnienia obserwacji przy aktualnym modelu)
  2. **E**: uaktualnienie modelu biorąc pod uwagę poprzedni etap
- Zaczyna od jakiegoś, wylosowanego modelu, dość jednorodnego, ale bez przesady.

Faza M

- Wykonujemy algorytm FB (mamy α, β i γ)
- Obliczamy: $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \mu)$
  - Czyli takie rozmyte wyjaśnienie historii
- (wzorki na poprzednim wykładzie)

Faza E. Obliczenia współczynników A

- Obliczamy $\hat{a}_{ij}$ jako:

$$\frac{\text{oczekiwana liczba przejść ze stanu i do stanu j}}{\text{oczekiwana liczba przejść ze stanu i}}$$

- Czyli:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,k)}$$

  gdzie $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \mu)$
- Zauważmy, że zwykłe wyliczenia $P(w_i w_j | w_i)$ robimy podobnie!

Faze E. Obliczenia współczynników B

Uwaga

Ponieważ interesuje nas tagowanie, będziemy liczyć $b_i(k)$, czyli prawdopodobieństwo emisji symbolu k pod warunkiem, że jesteśmy w stanie i (a nie pod warunkiem przejścia między stanami i a j.

Obliczamy b_{ik} jako:

$$\frac{\text{oczekiwana liczba przejść ze stanu i z emisją } o_k}{\text{oczekiwana liczba przejść ze stanu i}}$$

Obliczenia współczynników B (2)

- Wcześniej policzyliśmy prawdopodobieństwa bycia w stanie i: $\gamma_i(t) = P(X_t = i|O, \mu)$
- Rozpatrujemy te momenty, w których rzeczywiście wyemitowany został symbol o_k:

$$b_{ik} = \frac{\sum_{t=1, O_t=o_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

Gramatyczny opis języka

- Rozpoczniemy teraz omawianie bardziej zaawansowanych mechanizmów opisu języka
- Rozpoczniemy od wariantów gramatyk bezkontekstowych i regularnych,
- pokażemy, jak do nich dodawać wiedzę z zaczerpniętą z korpusu,
- a następnie pokażemy inne mechanizmy rozbioru

Zaczniemy od intuicyjnego poziomu opisu języka.

Noam Chomsky i gramatyki bezkontekstowe

- GBK (które zdefiniujemy formalnie za jakiś czas) miały w założeniu opisywać język naturalny (a nie C czy Pascala).
- Pierwsze (historyczne) drzewo rozbioru dla GBK pochodzi z pracy Noama Chomsky'ego z 1956 roku.
- Zdanie, które było „rozebrane" to:

  The man took the book.

  (rozbiór na tablicy)
- W węzłach wewnętrznych mamy nazwy kategorii gramatycznych, liśćmi są wyrazy angielskie.

Związki frazeologiczne

- Wyrazy łączą się w większe całości.
- Całości te mogą potem w zdaniu pełnić funkcję taką, jak pojedyncze wyrazy.

Uwaga

Jest to fundamentalna własność języka, używana w gramatycznych jego opisach.

Frazy nominalne

Fraza nominalna

Fraza nominalna to ciąg wyrazów, który w zdaniu pełni taką samą funkcję, jak pojedynczy rzeczownik.

Frazy nominalne odpowiadają pojęciom, więc są szczególnie istotne przy analizie języka naturalnego.
Przykłady:

- potęga zbudowana na zimnym makaronie
- kontrowersje wokół sztangistów z Korei Północnej
- telewizory plazmowe z Japonii,
- nowiutkie modele mercedesa

Frazy nominalne

Na poprzednim slajdzie przedstawione są **maksymalne** frazy. Ich podfrazy też mogą być frazami nominalnymi. Przykładowo

> *Koreańscy medaliści mistrzostw świata w podnoszeniu ciężarów*

Podfrazy

- podnoszeniu ciężarów
- mistrzostw świata
- Koreańscy medaliści
- medaliści mistrzostw świata
- mistrzostw świata w podnoszeniu ciężarów
- ...

Powyższe frazy na siebie nachodzą. Oczywiście rozbiór wybierze jeden z wariantów.

Prosta gramatyka bezkontekstowa

Opisujemy zdanie (zdania), takie jak:

The man took the book

Reguły gramatyki

sentence $\rightarrow$ NP VP
NP $\rightarrow$ Article Noun
VP $\rightarrow$ Verb NP
Verb $\rightarrow$ 'took' | 'read' | 'wrote' | ...
Article $\rightarrow$ 'a' | 'the'
Noun $\rightarrow$ 'man' | 'woman' | 'book' | 'newspaper' | ...

Generowanie lub analizowanie

- Gramatyka może działać jako generator lub jako analizator.
- W której z tych ról gramatyka z poprzedniego slajdu sprawi się lepiej?

Zdania:

The book read the man.
The man wrote the woman.
The woman took the woman.

Pytanie: Dlaczego analogiczna gramatyka dla języka polskiego byłaby nieco lepsza?
Odpowiedź: mielibyśmy osobą kategorię na podmiot i na dopełnienie (bo mają inne przypadki)

Związek zgody

- Podstawowe połączenie tworzące frazę nominalną to połączenie przymiotnika i rzeczownika.
- Przykładowo: piękna dolina, trudne twierdzenie, solidne podłoże.
- Inne przykłady: pięknej dolinie, trudnych twierdzeń, solidnego podłoża.
- Przymiotnik i rzeczownik muszą być w związku zgody, czyli mieć tę samą liczbę, przypadek i rodzaj.

Związki zgody w języku angielskim

Musi się zgadzać:

- liczba: flight leaves vs flights leave
- W pewnym sensie musi zgadzać się osoba

Frazy nominalne (2)

- Popatrzmy na następujące frazy nominalne: nowa klatka schodowa, dwie piękne i mądre dziewczyny, morderstwo w Orient Ekspesie, Partia Emetytów i Rencistów
- Widzimy następujące sposoby konstruowania fraz:
  - połączenie rzeczownika z przymiotnikiem (biały żagiel, klatka schodowa)
  - połączenie rzeczownika w wyrażeniem przyimkowym (np. dom na wzgórzu, niechęć do naleśników)
  - połączenie rzeczownika z rzeczownikiem w dopełniaczu (np. dom starców, synteza wodoru)
  - połączenie rzeczownika z liczebnikiem (dwaj mężczyźni, czternaście województw)
  - połączenie rzeczownika z rzeczownikiem za pomocą spójnika (Jaś i Małgosia)

Używanie spójników

Użycie spójników dotyczy nie tylko fraz nominalnych. Przykładowo:

Westchnął i usiadł na podłodze.
Woda lała się z prawa i z lewa.
Zrobił to szybko i z sensem.
W saturatorze można było kupić wodę z sokiem lub bez.
Szukał ołówka nad i pod biurkiem.

Frazy nominalne (3)

Frazy mogą powstawać przez wielokrotne użycie w.w. konstrukcji. Przykładowo

- stary dom na pobliskim wzgórzu
- krajowa partia emerytów i rencistów
- bliski znajomy poprzedniego ministra spraw wewnętrznych republiki górnej wolty
- słoń, którego trąba nigdy nie trafiła na kaktusa

Ostatni przykład pokazuje, że nie jest tak pięknie i aby w pełni opisywać frazy nominalne musimy de-facto opisać całą gramatykę (bo zdanie może być częścią frazy nominalnej).

Przykładowy formalizm opisu fraz (1)

```
np(L,P,R) ==> adj(L,P,R), np(L,P,R).
np(L,P,R) ==> np(L,P,R), adj(L,P,R).
np(L,P,R) ==> np(L,P,R), np(_,gen,_).
np(L,P,R) ==> np(L,P,R), prep(P2), np(_,P2,_).

np(pl,P,R1) ==> np(_,P,R1), [i], np(_,P,_R2).
                          %przybliżenie!
```

Jak łatwo zauważyć nawiązuje to trochę do prologu (w rzeczywistości jest programem prologowym)

Przykładowy formalizm opisu fraz (2)

Należy połączyć taką gramatykę ze słownikiem:

```
adv  ==> [X], {hasTag(X,adv:_)}.
prep(P) ==> [X], {hasTag(X, prep:P)}.
roman ==> [X], {roman(X)}.
nr ==> [X], {isNumber(X)}.

subst(L,P,R) ==> [X], {hasTag(X,subst:R:L:P)}.
```

Morfeusz w Prologu

```
tags(i, conj).
tags(w, prep:loc).
tags(bardzo, adv:_).

sot(1, [subst:sg:nom:f]).
sot(2, [subst:pl:nom:f, subst:sg:gen:f]).
sot(3, [adj:sg:nom:n, adj:pl:nom:f]).
sot(4, [adj:sg:gen:f, adj:sg:dat:f]).
sot(5, [adj:sg:nom:f]).

hasTag(Word, Tag) :- tags(Word, Tag).
hasTag(Word, Tag) :- ts(Word, SOT),
     sot(SOT, Tags), member(Tag, Tags).
```

Morfeusz w Prologu

```
ts(dziewczyna, 1).
ts(kobieta, 1).

ts(dziewczyny, 2).
ts(kobiety, 2).

ts(urodziwe, 3).
ts(inteligentne, 3).

ts(urodziwej, 4).
ts(inteligentnej, 4).
ts(nietrywialnej, 4).
```

Definite Clause Grammar

- Dlaczego używamy ==> zamiast -->?
- DCG nie lubi gramatyk lewostronnie rekurencyjnych, czyli takich:

$$X \rightarrow X\ Y\ X$$
$$X \rightarrow X\ Y$$

- A tę postać ma chociażby reguła dla spójnika `i`, czy dla połączenia rzeczownik – przymiotnik, we frazie liceum ekonomiczne.

Definicja gramatyki bezkontekstowej

Gramatyka bezkontekstowa jest definiowana przez 4 parametry: N, Σ, R, S.

- N jest skończonym zbiorem symboli nieterminalnych
- Σ, jest skończonym zbiorem symboli terminalnych, rozłączny z S.
- R jest skończonym zbiorem produkcji postaci:

$$A \to \beta$$

  gdzie $A \in N$, $\beta \in (\Sigma \cup N)^*$)
- S jest symbolem startowym, $S \in N$.

Wyprowadzenie w gramatyce bezkontekstowej

- Wyprowadzenie w jednym kroku:

$$\alpha A \gamma \Rightarrow \alpha \beta \gamma$$

jeżeli $A \to \beta \in R$

- Należenie do języka generowanego przez gramatykę:

$$w \in L(G) \text{ wtt } S \Rightarrow x_1 \Rightarrow x_2 \cdots \Rightarrow x_n = w$$

oraz $w \in \Sigma^*$ (przypominamy, że S to symbol startowy)

GBK a język naturalny

- Symbolami terminalnymi są słowa z języka.
- Symbolami nieterminalnymi są kategorie gramatyczne (być może z dodatkowymi parametrami)

Uwaga

Możemy dodawać symbolom nieterminalnym dodatkową strukturę, o ile liczba możliwych symboli pozostanie skończona. Stąd przykłady dla języka polskiego z symbolami np(L,P,R) nie wyprowadzają poza języki bezkontekstowe.

Podkategoryzacja

- Powiedzenie, że X jest czasownikiem (i podanie jego cech gramatycznych, takich jak liczba, czy rodzaj nie jest wystarczające, by orzekać o poprawności zdań z X.
- Popatrzmy na przykłady:

  Judyta dała Stefanowi czekoladki.
  Stefan nie myślał o Judycie, ale zabrał się bez wahania za czekoladki.
  Po tym wszystkim zatęsknił za kaszanką i zaczął płakać rzewnymi łzami

- Wszystkie czasowniki mają różne typy

Dygresja o typach

- Mówimy o typach nieformalnie, ale można nadać im znaczenie bardziej przypominające klasyczne typy (na przykład w językach programowania)
- Możemy wówczas myśleć na przykład o rzeczowniku ochota jako o funkcji, która bierze argument typu wyrażenie-przyimkowe-na-acc i zwraca pojęcie odpowiadające frazie nominalnej, na przykład ochota na naleśniki

Treebanks

Treebank

Bankiem drzew nazwiemy korpus zawierający drzewa rozbioru dla pojedynczych zdań.

- Dla języka angielskiego podstawowym korpusem jest Penn Treebank
- Dla języka polskiego podstawowym korpusem jest Składnica

O Składnicy

- Składnica powstała w dość specyficzny sposób:
  1. Wylosowane zostały pewne zdania z korpusu (20K)
  2. Użyto gramatyki **Świgra** (o niej za chwilę). Gramatyka ta generuje bardzo dużo rozbiorów
  3. W zdaniach, które się z sukcesem sparsowały, lingwiści wybierali właściwy rozbiór.
- Powstało w ten sposób 8227 zdań z rozbiorami.

Czy to jest dobra metoda?