N-gramowy model języka

Paweł Rychlikowski

Instytut Informatyki UWr

22 października 2018

Model językowy

Cel strategiczny 1

Chcemy umieć stwierdzić co jest, a co nie jest zdaniem w języku polskim (angielskim, chińskim)

Cel strategiczny 2

Należenie do języka chcemy trochę rozmyć, przypisując zdaniu:

X należy do języka polskiego

wartość rzeczywistą (na przykład z przedziału $[0, 1]$)

Uwaga

Model językowy może oceniać na przykład permutacje słów.

Modele N-gramowe. Przypomnienie

Definicja

N-gramem nazywamy ciąg kolejnych słów o długości N. 1-gramy to unigramy, 2-gramy to bigramy, 3-gramy to trigramy.

Za pomocą N-gramów tworzymy model języka, w którym staramy się przewidzieć kolejne słowo (N-te) na podstawie $N-1$ słów poprzednich.

- Czasem znajomość niewielkiego prefiksu wystarcza, by dobrze przewidywać kolejne słowo: dzisiaj mamy piękną **pogodę** (4-gramy)
- Często jest to trudne: Rzeźba przedplecza wyraźnie pomarszczona z kilkoma rozproszonymi punktami ku tylnym **odnóżom? rowkom? kawałkom? kątom? pokrywom? płaszczyznom?**

Lingwistyka korpusowa

- Obserwujemy przejście pomiędzy poniższymi sposobami uprawiania NLP:
  I. Ubieraniem intuicji lingwistycznych w precyzyjne i skomplikowane systemy opisujące język
  II. Badaniem prawidłowości obserwowanych w dużych (naprawdę!) zbiorach tekstów.
- Ale regułowe NLP, choć mniej modne, ciągle ma swoje miejsce!
- Powiemy sobie kiedyś więcej o pracy (Manning, 2011):

  Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?

Dostępne korpusy dla języka polskiego

Najważniejsze są dla nas 3 korpusy:

Narodowy Korpus Języka Polskiego

Dostępny w postaci nieoczyszczonych *N*-gramów (google: nkjp ngrams, 250M słów)

PolEval 2018

Korpus do konkursu PolEval 2018 (20M zdań, 451M słów)

Wikipedia

Dostępna pod adresem:
https://dumps.wikimedia.org/plwiki/latest/

NKJP nie ma części wspólnej z Wikipedią!

Prawdopodobieństwo sekwencji słów (powtórzone)

Prawdopodobieństwo sekwencji słów można obliczyć następująco:

$$P(w_1 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \dots P(w_n|w_1 \dots w_{n-1})$$

„Dalsze" prawdopodobieństwa szacujemy patrząc nie na całą historię, lecz na $N-1$ wyrazów poprzedzających wyraz przewidywany. Przykładowo dla $N = 2$ mamy

$$P(w_1 \dots w_n) \approx P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3) \dots P(w_n|w_{n-1})$$

Zdanie bardziej prawdopodobne = zdanie bardziej naturalne

Szacowanie prawdopodobieństw

Przykładowo dla bigramów mamy:

- $P(w_2|w_1) = \frac{P(w_1 w_2)}{P(w_1)}$.
- $P(w_1 w_2) = \frac{\mathrm{cnt}(w_1 w_2)}{N-1}$.
- $P(w_1) = \frac{\mathrm{cnt}(w_1)}{N}$.

Oczywiście przyjmujemy, że $N \approx N - 1$, co upraszcza wzory.

Szacowanie prawdopodobieństw

Przykładowo dla bigramów mamy:

- $P(w_2|w_1) = \frac{P(w_1 w_2)}{P(w_1)} = \frac{\mathrm{cnt}(w_1 w_2)}{\mathrm{cnt}(w_1)}$.
- $P(w_1 w_2) = \frac{\mathrm{cnt}(w_1 w_2)}{N-1}$.
- $P(w_1) = \frac{\mathrm{cnt}(w_1)}{N}$.

Oczywiście przyjmujemy, że $N \approx N - 1$, co upraszcza wzory.

Kilka uwag o przechowywaniu N-gramów w pamięci

N-gramy zajmują dużo miejsca

- **97%** 5-gramów z NKJP występuje dokładnie raz!

Jak je przechowywać (i inne proste uwagi po pracowni):

1. W Pythonie 3 napisy są kodowane w UTF-16 lub UTF-32, bytearray('żółć1234','utf8')
2. Warto zaprzyjaźnić się z sys.getsizeof(object)
3. Pusta struktura (napis, lista) zajmuje całkiem sporo miejsca:
   - bytearray: 33 bajty
   - Napis: 49 bajtów
   - Lista: 64 bajty
   - Słownik: 288 bajtów

O rozwiązaniach wykorzystujących dysk powiemy sobie jeszcze (m.in. na ćwiczeniach)

LM jako generator

- Prosty sposób na „organoleptyczne” sprawdzenie jakości modelu.
- Jeżeli zdania wydają się sensowne, to znaczy że można oczekiwać, że model sprawdzi się dobrze w innych zadaniach.

Generacja tekstów w modelach N-gramowych

Dla trigramów (zadanie 2 z listy):

1. Losujemy parę słów, od których zaczniemy generację
2. Sprawdzamy, jakie są możliwe kontynuacje tej pary
3. Losujemy z nich jedno słowo (z odpowiednim rozkładem)
4. Przesuwamy okno, czyli parą słów będzie teraz słowo nr 2 i słowo nr 3.

Uwaga

1. Zaczynamy od <BOS>
2. Kończymy na <EOS> (chyba, że nie wczytaliśmy wszystkich N-gramów)
3. Nie jest możliwa (ani sensowna) kontynuacja po EOS.

Odtwarzanie korpusu NKJP

- Będziemy używać 5-gramów.
- Dla zmniejszenia zużycia pamięci, pominiemy kwestię pamiętania rozkładu, czyli dla każdej czwórki pamiętamy listę możliwych kontynuacji.
- Nie przejmujemy się interpunkcją, bo ona w tym zadaniu akurat pomaga (dlaczego?)

Uwaga

Don't do it at home! :) – zużycie pamięci operacyjne przy naiwnej implementacji w Pythonie przekracza 50 GB.

Przykładowy fragment na następnym slajdzie. Czy umiesz podzielić „narrację" na części?

nie możemy firmować ani unii wolności, ani akcji wyborczej solidarność. panie premierze, nie krytykuję pańskiej osoby, tylko chcemy wyjaśnić - jako komisja - jakie były związki innych ludzi również związanych z polityką a działaniami chociażby personalnymi. otóż pan chyba dobrze wiedział, że i tak swych pieniędzy nie odzyska. skupowanie złotego piasku było nielegalne, gdyby więc trafił na trop oszustów, nie mógłby dochodzić swojej krzywdy. ci zresztą znikali zawsze bez śladu. prasa snuła przypuszczenia, że główne centrum handlarzy złotym piaskiem mieściło się w rydze. stamtąd nadchodziły bowiem listy z ofertami do ludzi, których miejscowi wspólnicy wskazywali jako potencjalnych nabywców. w listach zalecano, we wspólnym interesie, tajemnicę i wyznaczano spotkanie w warszawie bądź łodzi. gdy dochodziło do spotkania, sprzedający przynosili ze sobą charakterystyczne zamszowe woreczki, w jakie zazwyczaj pakowano prawdziwy złoty piasek. złoto w tej postaci stanowiło tylko zewnętrzną warstwę, natomiast reszta - to były miedziane opiłki. podobne operacje, przeważnie już na mniejszą skalę, odnosili także: jacek kazimierski (olympiakos pireus, kaa gent), józef wandzik (panathinaikos), adam matysek (bayer leverkusen), jacek bąk (olympique lyon), ...

nie możemy firmować ani unii wolności, ani akcji wyborczej solidarność. panie premierze, nie krytykuję pańskiej osoby, tylko chcemy wyjaśnić - jako komisja - jakie były związki innych ludzi również związanych z polityką a działaniami chociażby personalnymi. otóż pan chyba dobrze wiedział, że i tak swych pieniędzy nie odzyska. skupowanie złotego piasku było nielegalne, gdyby więc trafił na trop oszustów, nie mógłby dochodzić swojej krzywdy. ci zresztą znikali zawsze bez śladu. prasa snuła przypuszczenia, że główne centrum handlarzy złotym piaskiem mieściło się w rydze. stamtąd nadchodziły bowiem listy z ofertami do ludzi, których miejscowi wspólnicy wskazywali jako potencjalnych nabywców. w listach zalecano, we wspólnym interesie, tajemnicę i wyznaczano spotkanie w warszawie bądź łodzi. gdy dochodziło do spotkania, sprzedający przynosili ze sobą charakterystyczne zamszowe woreczki, w jakie zazwyczaj pakowano prawdziwy złoty piasek. złoto w tej postaci stanowiło tylko zewnętrzną warstwę, natomiast reszta - to były miedziane opiłki. podobne operacje, przeważnie już na mniejszą skalę, odnosili także: jacek kazimierski (olympiakos pireus, kaa gent), józef wandzik (panathinaikos), adam matysek (bayer leverkusen), jacek bąk (olympique lyon), ...

Zadanie z układaniem słów

- Rozwiązania heurystyczne: premiujemy jakoś fakt znalezienia w permutacji bigramów (3-gramów) z korpusu
  - (na przykład addytywna premia za każde wystąpienie w korpusie)
- Rozwiązanie kanoniczne: wybieramy permutację o większym prawdopodobieństwie w modelu.

Jak odróżnić Puszkina od Lermontowa?

- Tworzymy dwa LM, jeden nakarmiony dostępnymi dziełami Puszkina, drugi dziełami Lermontowa.
- Dla tekstu nieznanego autorstwa sprawdzamy jego prawdopodobieństwo w obu modelach.
- Uznajemy, że ten jest autorem, którego model zwraca większe prawdopodobieństwo

Problem!

Prawdopodobieństwa dla obu modeli najprawdopodobiej są równe 0.

Dlaczego?

Przypomnienie wzorów

Dla modelu bigramowego prawdopodobieństwo jest poniższym iloczynem:

$$P(w_1 \dots w_n) \approx P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3) \dots P(w_n|w_{n-1})$$

Oczywista uwaga

Jak gdzieś pojawi się wartość zerowa, wszystko inne przestaje mieć znaczenie!

Jak walczyć z zerami?

Istnieją dwie klasy metod pozwalających wyeliminować problem zerowych prawdopodobieństw (ktoś wie jakie?)

- „Zagęszczanie korpusu”, czyli łączenie słów w klasy.

  Nie ma w korpusie połączenia: unowocześniony helikopter, ale jest unowocześniony śmigłowiec oraz zmodernizowany helikopter.

- Modyfikacja prawdopodobieństw w taki sposób, żeby zawsze były niezerowe – tzn. by uznać (zgodnie z prawdą), że to, że czegoś nie obserwujemy, to nie oznacza, że to nie może się zdarzyć.

Class based N-Grams (1)

- Prostą metodą, używaną w systemach ASR tworzonych przez IBM jest tzw. IBM clustering.
- Założymy, że mamy podział wyrazów na klasy, z takimi klasami jak CITY, zwierającymi takie słowa jak Berlin, Paris, Warsow, Lisbona, Boston, ...
- Zakładamy, że każde słowo należy do dokładnie jednej klasy (hard clustering)
- Prawdopodobieństwa $P(w_i|w_{i-1})$ szacujemy w następujący sposób:

$$P(w_i|w_{i-1}) \approx P(c_i|c_{i-1}) \times P(w_i|c_i)$$

- Prawdopodobieństwa łatwo szacujemy z korpusu, przykładowo:

$$P(w|c) = \frac{C(w)}{C(c)}$$

Class based N-Grams (2)

- Inne przykładowe klasy: AIRLINE, DAYOFWEEK, MONTH, CITYNAME.
- Można zmodyfikować pojęcie słowa, tworząc takie pseudosłowa jak `New_York` czy `Sucha_Beskidzka`.
- Istnieją algorytmy automatycznego wyznaczania klastrów (na przykład Brown Clustering, używane przez IBM).
- Można łączyć N-gramy bazujące na klasach z innymi, wg wzoru:

$$P^*(w_i|w_{i-1}) = \alpha P(w_i|w_{i-1}) + (1-\alpha)P_{\text{class based}}(w_i|w_{i-1})$$

Jeszcze o klasach. Gramatyka

Uwaga

Ważnym i naturalnym źródłem podziału na klasy jest gramatyka danego języka.

Gramatycznym klasom słów poświęcimy więcej czasu, teraz przyjrzyjmy się plikowi supertags.txt i spróbujmy:

1. Odkryć zasadę, która stoi za tym plikiem.
2. Zastanowić się, jaki ma on związek z sufiksami.
3. Zastanowić się, jak może być użyteczny w zadaniu z układaniem słów.

Jeszcze o klasach. Lematy

Motywacja

Wykonujemy:

```
cat 2grams_cleaned | grep wiewiór | grep ' rud'
```

i otrzymujemy

```
2 ruda wiewiórka
1 rudej wiewiórki
1 wiewiórka ruda
2 rudy wiewiór
1 wiewiórkę rudą
1 wiewiórki rudymi
3 rude wiewiórki
```

Nie ma połączeń: rudym wiewiórkom, rudej wiewiórce, rudą wiewiórkę itd.

Jeszcze o klasach. Lematy (2)

Definicja 1

Paradygmatem odmiany danego słowa nazywamy zbiór wyrazów, stanowiących różne gramatycznie formy danego słowa (ups, jaka strasznie nieprecyzyjna definicja)

Definicja

Lematem słowa nazywamy wybraną zgodnie z ustaloną zasadą formę należącą do paradygmatu odmiany tego słowa.

Intuicja

Lemat jest tym, czego oczekujemy jako hasła w

1. słowniku języka polskiego, słowniku wyrazów obcych
2. słowniku polsko-angielskim

(w słowniku ortograficznym z kolei spodziewamy się pełnych paradygmatów odmiany)

Lematy i paradygmaty. Przykłady

- **czytanka**: czytanka czytankami czytanki czytance czytankach czytankę czytanko czytanek czytanką czytankom
- **czytać**: czytać czytałeś czytałybyście czytałaś czytali czytałabym czytał czytaj czytałoś czytałom czytano czytam czytała czytaliśmy czytałobyś czytaliby czytało czytałby czytałobym czytały czytałbyś czytałyście czytają czytałabyś czytałaby czyta czytajmy czytałyśmy czytać czytałybyśmy czytalibyśmy czytajmyż czytaliście czytamy czytałoby czytajże czytałam czytajcież czytacie czytalibyście czytajcie czytałem czytałbym czytasz czytałyby
- **obok**: obok

Jeszcze o lematach

Pytanie

Czy łatwo znajdziemy lemat dla wystąpienia słowa w tekście?

Trudne zdanie

Nie mam mam, że gdy płyną barki, to drżą kolana i barki.
Nie nabieraj matek, że gdy płyną okręty, to drżą kolana i ramiona

Nawiasem mówiąc nie też jest niejednoznaczna:

- jestem na nie! (**nie**)
- jestem na nie zły! (**on**)

Jest proste rozwiązanie tego problemu i trochę trudniejsze (ćwiczenia i kolejne wykłady).

Wygładzanie

Definicja

Wygładzanie (smoothing) polega na lepszym szacowaniu prawdopodobieństw rzeczy rzadko obserwowalnych w korpusie.

- Najprostszą formą wygładzania jest wygładzanie Laplace'a (+1).
- Uznajemy, że wszystko widzieliśmy o 1 raz za mało.

Wygładzanie Laplace'a

Uwaga

Zakładamy, że mamy słownik wszystkich słów danego języka, o wielkości V (czy to realistyczne założenie?)

- Dla unigramów, częstość wynosi:

$$P(w) = \frac{C(w) + 1}{N + V}$$

- Dla bigramów:

$$P(w_2|w1) = \frac{C(w_1 w_2) + 1}{C(w_1) + V}$$

- Zamiast 1 można dać λ, a zamiast V dać λV

Jak policzyć to, czego nie ma

- Chcielibyśmy mieć jakieś narzędzie, pozwalające tworzyć procedury wygładzania dostosowane do danych (a nie ad-hoc).
  - Na przykład, jak dostosować do danych λ z poprzedniego slajdu
- Oczywiście to, czego nie ma, ma częstość 0, czyli nie możemy przyjąć, że p-stwo to liczba wystąpień przez wielkość korpusu.
- Ale jeżeli podzielimy korpus na dwie części (K_1 i K_2), to jednej możemy użyć do liczenia prawdopodobieństw, a drugiej do szacowania, jakie są prawdopodobieństwa rzeczy niewidzianych.

Jak policzyć to, czego nie ma (2)

- K_1 powinien być większy, nawet dużo od K_2
- Na chwilę udajemy, że K_1 jest naszym całym korpusem, a zatem ten drugi korpus (K_2) to coś jakby spoza korpusu.
- Prawdopodobieństwo zobaczenienia nieznanego słowa można oszacować tak:

$$P(UNK) = \frac{\text{liczba słów z } K_2 \text{ , których nie ma w } K_1}{|K_2|}$$

- (oczywiście p-stwo konkretnego słowa otrzymujemy dzieląc $P(UNK)$ przez spodziewaną liczbę nieznanych słów)

Interpolacja

Innym pomysłem jest intepolacja, w której prawdopodobieństwo (np) trigramowe szacujemy wykorzystując również prawdopodobieństwo bigramów i unigramów.

$$P^*(w_3|w_1 w_2) = \lambda_1 * P(w_3) + \lambda_2 * P(w_3|w_2) + \lambda_3 * P(w_3|w_1 w_2)$$

, gdzie $\lambda_1 + \lambda_2 + \lambda_3 = 1$ oraz $\lambda_i > 0$.

Pytanie

Jak wyznaczyć wartości λ_i?

Deleted interpolation

Idea algorytmu jest następująca

- Przeglądamy wszystkie trigramy z K_2
- Dla każdego sprawdzamy, który sposób obliczenia prawdopodobieństwa jest najlepszy (tzn. daje największe prawdopodobieństwo):
  a) Trigramowo: $\frac{C(w_1 w_2 w_3)}{C(w_1 w_2)}$
  b) Bigramowo: $\frac{C(w_2 w_3)}{C(w_2)}$
  c) Unigramowo: $\frac{C(w_3)}{|K_1|}$

  (wszystkie częstości z K_1)
- Najlepszy sposób dostaje tyle punktów, ile razy dany trigram był w K_2
- Punkty normalizujemy, by sumowały się do jedynki.

W ten sposób staramy się zmaksymalizować prawdopodobieństwo korpusu.

Deleted interpolation w praktyce

- Wykonamy algorytm dla korpusu trigramowego (oraz dla bigramówi i unigramów) z NKJP
- Oprócz wartości, chcemy zobaczyć, jakie trigramy głosowały za różnymi sposobami liczenia prawdopodobieństwa.

Uwaga

Spróbujmy zastanowić się, jakich połączeń spodziewamy się w przypadkach tri-, bi- i unigramowego liczenia prawdodpodobieństwa

Wyniki procedury Deleted interpolation

- Otrzymaliśmy następujące wartości $\lambda_1 = 0.03$, $\lambda_2 = 0.11$, $\lambda_3 = 0.86$
- Połączenia:
  - Trigramowe: **za poczatkujacych rysowników**; **und regionale strukturpolityk**; **żyrafa zwiała kanałem**
  - Bigramowe: bardzo, **panie marszałku**; lub **suma ubezpieczenia**; wolności. **panie pośle**; w **'nicości absolutnej'**; narzędzie **rectangle (prostokąt)**
  - Unigramowe: przychodzi taki **i**; przyciąga uwagę **nie**; przycisku obecności. **bardzo**