Model probabilistyczny fleksji języka polskiego

Wojciech Jaworski

22 grudnia 2016

Zakładamy, że język jest rozkładem probabilistycznym na czwórkach (*form*,*lemma*,*cat*,*interp*), czyli, że wystąpienia kolejnych słów w tekście są od siebie niezależne. Interpretacja *interp* jest zbiorem tagów zgodnym a tagsetem SGJP. Kategoria $cat \in \{\text{noun}, \text{adj}, \text{adv}, \text{verb}, \text{other}\}$ Zakładamy też, że język jest poprawny, tzn. nie ma literówek, ani błędów gramatycznych.

Dysponujemy następującymi danymi:

- słownikiem gramatycznym S, czyli zbiorem czwórek, o których wiemy, że należą do języka;
- zbiorem reguł, czyli zbiorem czwórek (*fsuf*,*lsuf*,*cat*,*interp*)
- zbiorem wyjątków, czyli zbiorem czwórek, o których wiemy, że należą do języka, które nie są opisywane przez reguły
- otagowaną listą frekwencyjną.

Reguła przyłożona do formy ucina *fsuf* i przykleja *lsuf*.

Lista frekwencyjna wytworzona jest na podstawie NKJP1M. Usunięte zostały z niej symbole (formy do których odczytania nie wystarczy znajomość reguł wymowy takie, jak liczby zapisane cyframi, oznaczenia godzin i lat, znaki interpunkcyjne, skróty, emotikony). Usunięte zostały również formy odmienialne z użyciem myślnika i apostrofu (np. odmienione akronimy i nazwiska obce, formy takie jak „12-latek“). Interpretacje na liście frekwencyjnej zostały skonwertowane do postaci takiej jaka występuje w SGJP, łączącej interpretacje form identycznych. Na przykład interpretacje adj:pl:nom:m1:pos, adj:pl:voc:m1:pos, adj:pl:nom:p1:pos i adj:pl:voc:p1:pos zostały złączone w adj:pl:nom.voc:m1.p1:pos, a frekwencje form zsumowane.

Celem jest aproksymacja wartości P(*lemma*,*cat*,*interp*|*form*).

Pierwszym kryterium jest przynależność formy do słownika S. Jeśli forma należy do S zakładamy, że jedno z haseł S zawierające tę formę poprawnie opisuje jej lemat, kategorię i interpretację.

Zadanie 1. *Jakie jest prawdopodobieństwo trafienia na formę, której lemat, kategoria i interpretacja należy do słownika, czyli*

$$P((form, lemma, cat, interp) \in S)$$

Jakie jest prawdopodobieństwo trafienia na formę, która należy do słownika, ale jej lemat, kategoria lub interpretacja należy do słownika, czyli

$$P((form, lemma, cat, interp) \notin S \wedge form \in S)$$

Odpowiedź 1. *Prawdopodobieństwo natrafienia na formę należącą do słownika wynosi 95,67%, zaś natrafienia na formę należącą do SGJP bez odpowiedniej interpretacji – 3,92% (lista tych form znajduje się w pliku traps.txt).*

W przypadku form należących do słownika różnorodność interpretacji będzie niewielka, natomiast istotne będzie prawdopodobieństwo wystąpienia danego lematu. Zaś w przypadku form nie należących do słownika prawdopodobieństwo wystąpienia lematu będzie zawsze małe.

Dzielimy teraz listę frekwencyjną na część należącą do S i nie należącą do S. Od tej pory budujemy model osobno dla każdej z części.

W przypadku cześci należącej do S zauważamy, że

$$P(lemma, cat, interp|form) = P(form|lemma, cat, interp)\frac{P(lemma, cat, interp)}{P(form)}$$

Zakładamy, że *interp* jest niezależne od *lemma*, pod warunkiem określonego *cat*

$$P(lemma, cat, interp) = P(lemma, cat)P(interp|lemma, cat) = P(lemma, cat)P(interp|cat)$$

$P(form)$, $P(lemma, cat)$ i $P(interp|cat)$ szacujemy na podstawie listy frekwencyjnej, w przypadku pierwszych dwu stosując wygładzanie. Wyliczenie $P(form)$ zawiera uogólniona lista frekwencyjna (ścieżka `resources/NKJP1M/NKJP1M-generalized-frequency.tab` w repozytorium ENIAM), $P(lemma, cat)$ – plik `prob_lemmacat.txt`, zaś $P(interp|cat)$ – `prob_itp_givencat.txt` (oba zawarte w katalogu `morphology/doc`).

$P(form|lemma, cat, interp)$ wynosi 0, gdy w S nie ma krotki postaci (*form*,*lemma*,*cat*,*interp*); 1, gdy jest dokładnie jedna krotka z (*lemma*,*cat*,*interp*). Gdy jest ich więcej oznacza to, że lemat ma przynajmniej dwa warianty odmiany. Są to przypadki rzadkie. Przypisujemy każdej z możliwości prawdopodobieństwo 1.

Zadanie 2. *Przejrzeć SGJP i znaleźć wszystkie przykłady, w których dla ustalonego lematu, kategorii i interpretacji jest więcej niż jedna forma. Znaleźć wystąpienia tych krotek na liście frekwencyjnej.*

Odpowiedź 2. *Lista takich form znajduje się w pliku multi_forms.txt.*

Teraz zanalizujemy drugą część listy frekwencyjnej. Załóżmy, że reguły mają postać taką, że sufiks żadnej reguły nie jest podciągiem sufixu innej z nich.

Sufiksy reguł tworzą drzewo, które w każdym węźle ma dowiązania do sufixów o jeden znak dłuższych oraz kategorię pozostałe traktową łącznie. Przyjmujemy następujące założenie modelowe:

$$P(lemma, cat, interp|form) \approx P(rule|form) = P(rule|fsuf)$$

Wynika ono z tego, że mając nieznaną formę musimy oprzeć się na ogólnych regułach odmiany i nie możemy korzystać z tego że ma ona jakieś konkretne brzmienie. Korzystamy tutaj tylko z reguł oznaczonych jako produktywne.

Problem tu jest taki, że lista frekwencyjna jest zbyt mała by precyzyjnie określić p-stwo ok. 40000 reguł. Dlatego znowu stosujemy zabieg z prawdopodobieństwem warunkowym.

$$P(rule|fsuf) = P(lsuf, cat, interp|fsuf) = P(fsuf|lsuf, cat, interp)\frac{P(lsuf, cat, interp)}{P(fsuf)}$$

$P(fsuf)$ jest prawdopodobieństwem tego, że do języka należy słowo o zadanym sufixie. Można je oszacować za pomocą listy frekwencyjnej.

Zakładamy, że *interp* jest niezależne od *lsuf*, pod warunkiem określonego *cat*

$$P(lsuf, cat, interp) = P(lsuf, cat)P(interp|lsuf, cat) = P(lsuf, cat)P(interp|cat)$$

$P(lsuf, cat)$ i $P(interp|cat)$ można oszacować na podstawie listy frekwencyjnej.

Zadanie 3. *Oszacować $P(fsuf)$ i $P(lsuf, cat)$ na podstawie listy frekwencyjnej. Sprawdzić dla jakich sufiksów próbka jest mała albo nie ma jej wcale.*

$P(fsuf|lsuf, cat, interp)$ wynosi 0, gdy nie ma reguły postaci (*fsuf*,*lsuf*,*cat*,*interp*); 1, gdy jest dokładnie jedna reguła z (*fsuf*,*lsuf*,*cat*,*interp*). Ustawiamy produktywność reguł tak by nie pojawiało się więcej pasujących reguł.

Zadanie 4. *Określić produktywność reguł i sprawdzić, czy nie ma niejednoznacznych dopasowań.*

Zadanie 5. *Określić jakość modelu.*

Odpowiedź 3. *Wyliczona jakość modelu (stopień pokrycia listy frekwencyjnej przez co najmniej 95% najbardziej prawdopodobnych interpretacji wg modelu) wyniosła 79,90%.*

Pytanie 4: Czy powyższe przybliżenie jest poprawne, jak często jest więcej niż jedna reguła i ile wynoszą wówczas p-stwa?

Zadania poboczne: wytworzenie otagowanej listy frekwencyjnej, wytworzenie (uzupełnienie) zbioru reguł na podstawie SGJP i listy frekwencyjnej, wskazanie, które reguły opisują sytuacje wyjątkowe.

Zadanie na przyszłość: reguły słowotwórstwa i ich interpretacja semantyczna.

Do powyższego modelu trzeba jeszcze dodać prefixy nie i naj.