Przetwarzanie języka naturalnego. Wstęp

Paweł Rychlikowski

Instytut Informatyki UWr

11 października 2018

Przetwarzanie języka naturalnego

- Język naturalny vs. język sztuczny
- Przetwarzanie tekstów, nie języka
- Przykładowe zadania:
  - Tłumaczenie z polskiego na angielski
  - Streszczanie tekstów
  - Part of speech tagging

Przetwarzanie języka naturalnego

- Język naturalny vs. język sztuczny
- Przetwarzanie tekstów, nie języka
- Przykładowe zadania:
  - Tłumaczenie z polskiego na angielski - łatwa (?) definicja
  - Streszczanie tekstów – co to jest dobre streszczenie
  - Part of speech tagging – zadanie dość techniczne, pomocnicze

Jak język pojawia się w nauce

- Lingwistyka
- Psycholingwistyka
- Filozofia
- Lingwistyka obliczeniowa

Jak język pojawia się w nauce

- **Lingwistyka** posługuje się intuicją dobrego zdania
- **Psycholingwistyka** bada sposób używania i uczenia się języka
- **Filozofia języka**, bada relację między osobą, językiem i światem
- **Lingwistyka obliczeniowa** modeluje matematycznie/informatycznie język

NLP w informatyce

- Teoria języków formalnych i translatory (Noam Chomsky nie chciał tworzyć JFiZO, a opisywać język angielski)
- Data mining (text mining) – szukanie prawidłowości w danych
- Machine learning
- Sztuczna inteligencja

Problemy AI zupełne

AI completeness

Problem jest AI-zupełny, jeżeli do jego rozwiązania konieczne jest pełne modelowanie ludzkiej inteligencji.

Przykład: czy tekst jest koherentny?

John hid Bill's car keys. He was drunk
John hid Bill's car keys. He likes spinach

Która z wypowiedzi jest ok? Kto to jest he?

Jasiu je słonecznik i...

Rozważmy zdanie:

Jasiu je słonecznik i śmieci.

Słowo „śmieci” to może być:

1. liczba mnoga od „śmieć”
2. czasownik „śmiecić”

Ale: dzieci jedzą śmieci

Jak zrobić kaczkę?

Czy łatwo napisać program, który umie znaleźć wszystkie znaczenia zdania:

I made her duck.

5 znaczeń!

- To ja zrobiłem jej kaczkę (origami?)!
- Ugotowałem dla niej kaczkę (w pomarańczach, na przykład)
- Ugotowałem kaczkę, która należała do niej
- Zmieniłem ją w kaczkę (magiczną różdżką)
- Sprawiłem, że zrobiła unik.

NLP jest trudne

- AI – zupełność
- Test Turinga jako wyznacznik inteligencji
- Łatwe znajdywanie trudnych kontrprzykładów dla (pozornie?) łatwych zadań

Zadanie

Jak po polsku przeczytać słowo: „por”?

Zupa porucznika Kowalskiego

Popatrzmy na zdania:

> *Zawsze mówiłem temu Kowalskiemu:* ***należy włożyć do zupy por. Kowalskiego*** *to jednak nie przekonywało.*
>
> *Wraz z sierż. Nowakiem zastanawialiśmy się, co* ***należy włożyć do zupy por. Kowalskiego****, żeby była tak dobra jak ostatnio.*

Por. chociażby por kpt. Nowaka lub seler por. Kowalskiego.

NLP jest łatwe

- Niektóre zadania da się rozwiązywać z bardzo wysoką skutecznością prostymi metodami:
  1) Podział na zdania (dlaczego?)
  2) Czy „play" to czasownik, czy rzeczownik? (efektywność tagerów to około 97.5% dla j.ang.)
  3) Wyszukiwanie informacji
  4) Sukcesy niektórych botów konwersujących.
- Nasz bot poetwanna.be zdobył ex-equo pierwsze miejsce na konkursie NIPS 2017 Conversational Challenge.

Uwaga

O tym co z NLP jest przydatne w tworzenu bota będziemy jeszcze mówić!

Eliza

- Eliza udawała psychoterapeutę. Z wielkim sukcesem!
- Krótki (hipotetyczny) fragment reguł Elisy:

```
I am  -> YOU ARE
my -> your
YOU ARE (depressed|sad)  -> why do you think you are \1
.* always .* -> can you think of a specific example
```

Test Turinga i nagroda Loebnera

- Alan Turing zaproponował słynny test, którego przejście oznacza Stworzenie Sztucznej Inteligencji.
- Co roku odbywają się zawody pomiędzy botami konwersacyjnymi, których naturalność jest oceniana przez sędziów (zawody o nagrodę Loebnera)
- Program, który „przejdzie test", otrzyma „dużą nagrodę Loebnera"
- Zdarza się, że niektóre programy, w niektórych sytuacjach potrafią zmylić niektórych sędziów.

Uwaga dla zainteresowanych

Istnieje całkiem spory bot ALICE, z dostępnym kodem w (dość paskudnym) języku, jakim jest AIML. Są interpretery do AIML w Pythonie.

Przykładowe zadania NLP

1. Automatyczne tłumaczenie (z jednego języka na inny)
2. Generowanie streszczeń (jednego lub wielu dokumentów)
3. Rozpoznawanie mowy i systemy dyktowania
4. Interfejsy w języku naturalnym (na przykład telefoniczna rezerwacja biletów)
5. Odpowiadanie na pytania i wyszukiwanie informacji
6. Ocena nastawienia autora do opisywanego obiektu (czy opinia na forum jest pozytywna? – łatwe, ale sarkazm czy ironia problematyczne)
7. Ocena wiarygodności tekstu (na przykład wypowiedzi na forum)
8. Automatyczne poprawianie błędów (literówki, Grammarly)
9. ... i wiele innych

Język [C]hiński

Pytanie

Co łączy język C i język chiński?

1. Choć można na oba patrzeć jako na ciągi znaków...
2. to w analizie wygodnie jest wyodrębnić **tokeny**.
3. Nie możemy posiłkować się (tylko) spacjami (bo np. w chińskim ich nie ma)

Pytanie

Jak przeprowadzić tokenizację tekstu?

C vs. Chiński

- Referencyjny algorytm tokenizecji dla chińskiego to algorytm **MaxMatch** (taki sam jak dla C)
- Czyli: pierwszym tokenem tekstu jest najdłuższy jego prefiks, który jest zarazem poprawnym tokenem.
- W *bezspacjowym* angielskim działa źle (w polskim?)
  - wecanonlyseeashortdistanceahead

C vs. Chiński

- Referencyjny algorytm tokenizecji dla chińskiego to algorytm **MaxMatch** (taki sam jak dla C)
- Czyli: pierwszym tokenem tekstu jest najdłuższy jego prefiks, który jest zarazem poprawnym tokenem.
- W *bezspacjowym* angielskim działa źle (w polskim?)
  - we canon l y see ash ort distance ahead

Jeszcze o tokenach

- Język C ma bardzo ściśle zdefiniowane tokeny (i jednoznaczną tokenizację).
- Dla języków naturalnych nie ma tak dobrze:
  - niebiesko-czarni (1 czy 2 tokeny?)
  - Mi-24 (1 czy 3 tokeny?)
  - m.in. (1, 2, czy 4 tokeny?)

Pewne decyzje są arbitralne (i trzeba się z tym pogodzić, w miarę starając się zachować konsekwencję).

Niejednoznaczność tokenizacji?

Kropka po skrócie przyklejona, kropka jako znak końca zdania – osobny token. (patrz: problem pora i porucznika)

Jeszcze o tokenach (2)

Czasem pomija się tokenizację, traktując język np. jako:

- Ciąg znaków (ASCII, Unicode, Latin-2?)
- Ciąg bajtów (kodowanie utf-8)

Co to jest język polski?

Język (podobnie jak na JFiZO) można traktować jako zbiór wypowiedzi (ciągów słów-tokenów). Czyli *język polski* to **zbiór poprawnych zdań z języka polskiego**.

Ta prosta definicja jest nieco problematyczna. Popatrzmy na przykładowe zdania:

- Zielone bezbarwne idee drzemią wściekle.
- Fioletowa sytuacja podskoczyła wytrwale.
- Małe czarne dzieci śpią spokojnie.
- Mały czarny dziecko spać spokojnie.
- Mały czarną dzieciami śpicie spokojnego.
- Ja tu przyszłem bynajmniej do pani (z karnistrem!).

Co to jest język polski?

Moja odpowiedź (co jest, co nie jest zdaniem polskim).

- Zielone bezbarwne idee drzemią wściekle.
- Fioletowa sytuacja podskoczyła wytrwale.
- Małe czarne dzieci śpią spokojnie.
- Mały czarny dziecko spać spokojnie.
- Mały czarną dzieciami śpicie spokojnego.
- Ja tu przyszłem bynajmniej do pani (z karnistrem!).

Ale pewnie są inne odpowiedzi.

Język naturalny jako język formalny

- „Literami" są słowa (czy „alfabet" Σ jest skończony?)
- Czy język jest tworem skończonym? (czy zdania sa dowolnie długie?)
- Czy do opisu wystarczy prosty formalizm, na przykład język regularny (zastanówmy się!)

Uwaga

Alfabet (zbiór słów) można łatwo uczynić skończonym:

a) Dodając sztuczne słowo **OOV** (Out of Vocabulary)

b) Zamieniając rzadkie słowa na k-literowe sufiksy

c) Zamieniając pewne słowa na ich części mowy (że np. rzeczownik, przymiotnik, można też tworzyć inne klasy)

d) Bytes Pair Encoding, BPE

Bytes Pair Encoding

W wielkim skrócie:

a) Liczymy słowa w korpusie (czyli dużym, reprezentatywnym, zbiorze tekstów)

b) W słowach liczymy częstości par liter
   - Jeżeli abrakadabra występowało 15 razy, to zwiększamy licznik **ra** o 30.

c) Zamieniamy najczęstszą parę na nową (pseudo)literę

d) Czynności powtarzamy aż do otrzymania pożądanej liczby pseudoliter.

Każde słowo reprezentujemy jako ciąg pseudoliter (szczegóły na ćwiczeniach).

O słoniu, trąbie i zakończeniu

Język polski jako język formalny

Język regularny raczej nie wystarczy. Przyjrzyjmy się zdaniu:

> *Słoń, którego trąba, której zakończenie było wilgotne, była długa, kołysał się nerwowo.*

Obserwujemy sekwencje rodzajów: **mżnnżm**. Język słów postaci ww^R jest nieregularny.

Analiza słonia

Ładne fragmenty zdania o słoniu (głębokość 2):

- ...trąba, której zakończenie było wilgotne, była długa...
- Słoń, którego trąba była długa, kołysał się nerwowo.

Chyba wygodnie jest uznać, że zdanie o słoniu z głębokością 3 jest również prawidłowe.

Czy języki bezkontekstowe wystarczą?

- Odpowiedź nie jest oczywista, ale w niektórych językach wystąpują konstrukcje typu: $a_1 a_2 \ldots a_n b_1 b_2 \ldots b_n$
- Przykład (nienajlepszy...)

  Asię, Bartka i Celinę opisują najlepiej następujące cechy: mądra, kłótliwy, pyskata.

- Istnieją formalizmy gramatyczne (Gramatyki łagodnie kontekstowe), które są pomiędzy gramatykami bezkontekstowymi, a kontekstowymi, utworzone specjalnie dla języka naturalnego.
- Obecnie: raczej marginalne znaczenie (wyjaśnimy sobie to kiedyś dokładniej)

Model językowy

Model językowy (Language model, LM)

Model językowy definiuje dla zdania pewną wartość liczbową, która mówi o tym, jak bardzo naturalnym jest, że to zdanie należy do języka.

Do czego to może służyć? Przykładowe zastosowania:

1. Ocena wygenerowanej wypowiedzi (mnóstwo różnych zastosowań!)
2. Ocena poprawności tekstu (przy nauczaniu języka obcego)
3. Analiza autorstwa tekstu (różne modele jednego języka: przykład Puszkina i Lermontowa)

Modele N-gramowe

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.

Przykłady

- Dzisiaj mamy piękną ...
- Basia jest szczęśliwa, bo dostała w szkole ...
- Mam 11 lat i ...

Przykłady

- Dzisiaj mamy piękną pogodę
- Basia jest szczęśliwa, bo dostała w szkole szóstkę
- Mam 11 lat i już to zrobiłam (by Google suggest)

Analiza przykładów

- Wydaje się, że N powinno być duże; w przykładach wynosi 4, 8 i 5.
- Mniejsze fragmenty nie wystarczą w zdaniach 1 i 3:
  - mamy piękną ...
  - 11 lat i ...
- 5-gramy nie wystarczą do zdania z Basią:
  bo dostała w szkole ?
- N-gramy musimy brać z jakiegoś źródła. Jeżeli źródłem są pytania gimnazjalistów, to ...

Korpus

Definicja

Korpusem nazywamy duży zbiór tekstów, o którym zakładamy, że dostarcza nam wiedzy o języku.

Na podstawie korpusu będziemy szacować prawdopodobieństwa pewnych sekwencji, zliczając liczbę ich występień w korpusie.

Korpus (cd)

Korpusy mają kilka parametrów:

- Wiekość: Narodowy Korpus Języka Polskiego ma 250 mln. słów
- Dziedzina: korpus może być z jakiejś dziedziny (np. korpus medyczny, prawny, etc), ale nie musi.
- Sposób przetworzenia: korpusy mogą być gołym tekstem, mogą być podzielone na artykuły, ale mogą też zawierać różną ilość dodatkowych informacji.

Co można dodać do korpusu

- Nic (całkiem fajna opcja, nie wprowadza zniekształceń, wygodna dla programistów)
- Podział na tokeny (jak w językach programowania, elementarne składowe tekstu)
- Podział na zdania
- Zaznaczone frazy (czyli ciągi wyrazów o określonej semantyce, stanowiące całość)
- Drzewa rozbioru zdań

W praktyce dodatkowe informacje są często dodawane półautomatycznie, oczywiście najbardziej wartościowy jest korpus anotowany całkowicie ręcznie.

Reprezentatywność korpusu

Lingwiści dobierający teksty starają się, by korpus „ogólny” był reprezentatywny, czyli mówił coś o języku w całości. Co to może znaczyć?
Na przykład, że określamy jakoś proporcje różnych rodzajów tekstu i następnie ich się trzymamy. Tzw. korpus frekwencyjny języka polskiego zawierał mieszankę

- tekstów popularnonaukowych
- drobnych wiadomości prasowych
- tekstów publicystycznych
- prozy artystycznej
- dramatu artystycznego

Dostępne korpusy dla języka polskiego

- Frekwencyjny korpus języka polskiego (około 500 tys. słów)
- Narodowy Korpus Języka Polskiego nie jest dostępny w całości, ale dostępne są statystyki n-gramów (dla n=1,2,3,4,5), zobacz stronę nkjp.pl, ewentualnie pelcra.pl oraz korpus.pl
- Korpus Rzeczpospolitej
- Korpus Politechniki Wrocławskiej (KPWR, wiele dodatkowych informacji)
- Składnica (ok 6 tys. zdań wraz z rozbiorami).
- Wolne Lektury oraz WikiSources

Nowość

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

Inne korpusy

Istnieją kolekcje dokumentów, które mogą działać jako korpus:

- Wikipedia (również jako korpus wielojęzyczny)
- Akty prawne (zwłaszcza Unii Europejskiej, również wielojęzyczny)
- Normy i patenty
- Uzasadnienia orzeczeń sądowych, na przykład `http://orzeczenia.nsa.gov.pl`
- Zapisy posiedzeń sejmu (połączenie tekstów i audio)
- Artykuły z konkretnej gazety z wielu lat.

O pełności korpusu (bigramy)

- Dla 50000 słów mamy 2 miliardy 500 tys potencjalnych par słów.
- Nawet pary pozornie bez związku mogą występować w sensownych zdaniach:
  - walencyjny znacząco
  - ekonomicznego do
  - dwa babuleńka

O pełności korpusu (bigramy)

- Dla 50000 słów mamy 2 miliardy 500 tys potencjalnych par słów.
- Nawet pary pozornie bez związku mogą występować w sensownych zdaniach:
  - walencyjny znacząco (0 trafień by Google)
  - ekonomicznego do (ok. 9 tys trafień by Google)
  - dwa babuleńka (0 trafień)

Przykłady

- Słownik walencyjny znacząco poprawił komfort pracy nad językiem.
- Od II Liceum ekonomicznego do najbliższego pubu jest bardzo blisko
- Dwa babuleńka miała jedynie koziołki i biedę straszną klepała.

Pełność korpusu trigramowego (i więcej)

- Brak trafień w Google'u dla tak naturalnych fraz jak:

  studentka wstała energicznie
  zmęczony student przysnął
  wykładowca powiedział głośno

- Mimo, że nie wszystkie teksty są indeksowane przez Google, wydaje się, że inne korpusy są generalnie jeszcze bardziej dziurawe.

Łatanie korpusu (część 1)

- Niezerową liczbę trafień mają następujące trigramy:

  dziewczyna wstała energicznie
  zmęczony student zasnął
  wykładowca stwierdził stanowczo

Wykrzystanie korpusu

Celem jest wykorzystanie korpusu do utworzenia modelu językowego.

Pamiętajmy o pewnej dyscyplinie w wykorzystaniu danych:

- Nie można tych samych danych wykorzystywać do *uczenia* oraz *testowania*.
- Korpus powinien być dostosowany do zadania, które chcemy rozwiązać.