Zanurzeń ciąg dalszy, trochę gramatyki i POS-tagging

Paweł Rychlikowski

Instytut Informatyki UWr

20 listopada 2018

Zanurzenia słów z lotu ptaka

Dane wejściowe

Duży zbiór tekstów (korpus).
I nic więcej! (żadnych dodatkowych danych, typu klasy słów, etc).

Cel (pozorny?)

Obliczenia mają na celu stworzenie mechanizmu, który (w zależności od wariantu):

1. Dla pary słowo i kontekst przewiduje, czy pasują do siebie (o tym mówiliśmy)
2. Dla danego słowa w_i przewiduje niezbyt odległe słowo w_j (też word2vec, ale o tym powiemy kiedy indziej)
3. Dla słowa i kontekstu przewiduje PPMI (GloVe, konkurent word2vec, będzie trochę na ćwiczeniach).

Zanurzenia słów z lotu ptaka (2)

Wynik

1. Przypisanie słowo $\rightarrow$ wektor (dla słów)
2. Przypisanie słowo $\rightarrow$ wektor (dla kontekstów)

Możemy wziąć pierwsze, skleić lub dodać oba.

Cel podstawowy

Posiadanie wektorowej reprezentacji dla słów

Co jeszcze można zanurzać?

Wszystko! (prawie)

- Tytuły wikipediowe (tekst składa się ze zdań: artykuł link-wychodzący)
- Lematy słów
- Sufiksy słów, ewentualnie opisy gramatyczne
- Produkty i klientów, którzy je kupili (tworząc tym samym systemy rekomendujące)

Zanurzanie fraz

- „Pseudosłowa”, takie jak new_york, czarna_śmierć, czy czerwone_wino, można traktować jak zwykłe słowa i wyznaczać ich zanurzenia.
- Jeden z wariantów word2vec coś takiego robił, uznając, że fraza to coś, co ma odpowiednio wysoką wartość score:

$$\text{score}(w_i w_j) = \frac{\text{cnt}(w_i w_j) - \delta}{\text{cnt}(w_i)\text{cnt}(w_j)}$$

- Podobnie jak w BPE robimy więcej iteracji i dłuższe frazy (od 2 do 4 iteracji).

Zanurzanie fraz (2)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Źródło: Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality

Zanurzanie zdań

Uwaga

Ze zdaniami sytuacja jest inna: nie możemy ich traktować jako atomów.

Pomysł podstawowy

Suma zanurzeń wyrazów (może też fraz) z ewentualnymi wagami. Działa całkiem przyzwoicie!

Ocenianie zanurzeń

Sposób podstawowy

Korzystamy z ręcznie przygotowanych danych, w których ludzie oceniali, jak bardzo (np. w skali 0 do 10) są do siebie podobne dwa słowa (ewentualnie dwa zdania)

WordSim 353

love sex 6.77
tiger cat 7.35
book paper 7.46
television radio 6.77
smart stupid 5.81
company stock 7.08
professor cucumber 0.31
king cabbage 0.23
king queen 8.58
king rook 5.92
bishop rabbi 6.69
glass magician 2.08

Ocenianie zanurzeń (2)

Uwaga

Łączy dwa rodzaje podobieństwa: similarity oraz relatedness!
soap – opera vs king – queen

Inne uwagi:

- Istnieją podziały tego zbioru na części i wiele innych podobnych danych (na ćwiczeniach: Problems With Evaluation of Word Embeddings Using Word Similarity Tasks)
- Mieszają się znaczenia, stąd relacja podobieństwa jest nieprzechodnia:
  - Z tego że queen jest podobne do king i king jest podobne do rook nie wynika... (ok, to może nie najlepszy przykład :))

Jeszcze o grupach słów

- Mówiliśmy o grupach słów w kontekście zmniejszania liczby wymiarów dla gęstych wektorów.
- **Uwaga**: znajomość grup nie jest nam potrzebna (choć oczywiście moglibyśmy wyznaczyć zanurzenia dla grup, gdybyśmy wiedzieli, jakie słowa należą do grupy.
- Można sobie wyobrazić półautomatyczne tworzenie grup, nazwiemy ten algorytm (nieoficjalnie): **Algorytmem Elitarnego Klubu**

Algorytm Elitarnego Klubu

1. Ręcznie podajemy zbiór członków założycieli (czyli kilka słów z danej grupy)
   - Przykładowo: mucha, motyl, żuk, modliszka (owady)
2. Do zbioru słów włączamy jedynie te słowa, które są bardzo podobne (cosinusy!) do co najmniej K wyrazów ze zbioru
   - Czyli do klubu można wejść mając rekomendację (na przykład) 3 członków.

Powracając do zasad gramatyki

- Za pomocą części mowy chcemy opisywać język.
- Zamiast wymieniać wszystkie zdania, chcemy móc powiedzieć, że ciąg:
  `przymiotnik rzeczownik przysłowek czasownik przyimek rzeczownik` jest poprawnym gramatycznie zdaniem.
- Przykładowo:

  Szary wróbel zwinnie udiadł na drzewie.
  Sprytny lis okrutnie zadrwił z kruka.

- Ale z drugiej strony nie są zdaniami:

  Szaremu wróblem zwinnie usiedli na drzewu.
  Sprytni lisa okrutnie zadrwiła z krukowi.

Parametry morfosyntaktyczne (rzeczowniki)

- Rzeczowniki mają następujące parametry:
  - Liczba (pojedyncza, mnoga)
  - Przypadek (mianownik celownik, dopełniacz, biernik, narzędnik, miejscownik, wołacz)
  - Rodzaj (męski, żeński i nijaki)
- Jedno słowo może mieć więcej niż jeden zestaw tych parametrów:

  Dziewczyny *w naszej klasie są raczej fajne,*
  zatem nie ma w naszej klasie niefajnej ***dziewczyny****.*

- (mnoga, mianownik, żeński) vs (pojedyncza, dopełniacz, żeński)

Parametry morfosyntaktyczne (przymiotniki)

- Przymiotniki mają wszystkie parametry rzeczowników.
- Dodatkowo posiadają stopień. Przykładowe stopniowanie: równy, równiejszy i najrównejszy.
- Można też dodać informację, czy sa pozytywne (ładny), czy negatywne (nieładny).

Parametry morfosyntaktyczne (czasowniki)

- Sytuacja z czasownikami jest bardziej skomplikowana (i tu odchodzimy nieco od szkolnych klasyfikacji).
- Przede wszystkim dzielimy je na formy przeszłe i nieprzeszłe (czyta lub przeczyta vs przeczytał)
- Czas wydaje się nieużyteczną klasyfikacją (bo czyta i przeczyta, choć odnoszą się do teraźniejszości i przyszłości, zachowują się tak samo, a inaczej niż przeczytał
- Wyodrębniamy ponadto specjalne formy, takie jak:
  - bezokolicznik (czytać),
  - rozkaźnik (czytaj, czytajcie)
  - bezosobnik (czytano)

Parametry morfosyntaktyczne (czasowniki) (2)

- W formach przeszłych występują rodzaje, liczby i osoby: powiedział, powiedziała, powiedziało, powiedziałam, powiedzieliście, ...
- W formach nieprzeszłych występują liczby i osoby (czytam, czytasz, czyta, czytamy, czytacie, czytają), ale nie rodzaje
- W rozkaźniku mamy tylko liczby (czytaj, czytajcie)

Segment a słowo

- Niektóre korpusy i tagsety mówią o segmentach, a nie o słowach.
- powiedzilibyście = powiedzieli+by+ście, powiedzałam = powiedziała+m, powiedziałem = powiedział+em
- Uwaga: zauważamy, że wówczas formy przeszłe nie mają osób (bo na osobę wpływa doklejony segment, tzw. aglutynant -m, -ś, -śmy, -ście)
- Jeżeli przyjmiemy, że nie używamy aglutynantów, wówczas potrzebujemy jeszcze jednej kategorii, na tryb przypuszczający.

Po co to?

Pytanie

Czy rzeczowniki (przymiotniki, przysłówki) odmieniają się przez osoby?

Pięknieś to wyszykowała. On to zrobiłby dużo gorzej. Tak, on by to zrobił naprawdę dużo gorzej.Alem dzisiaj domem się zajęła, choć do parkum pójść mogła.

- Wydaje się, że problem, który rozwiązujemy istnieje, choć dotyczy raczej wyrazów archaicznych i jest mało istotny w zwykłych tekstach.

Paradygmat odmiany. Przypomnienie (częściowo)

- Słowa mają pewne właściwości (na przykład krowa ma rodzaj żeński)
- Słowa mogą się odmieniać przez właściwości (na przykład krowa odmienia się przez liczby i przypadki, i mamy formy: krowa, krową, krowami, krowach, itd).
- Dla każdego słowa możemy wyznaczyć jego formę bazową (tzw. lemat)

Paradygmat odmiany

Zbiór słów wszystkich słow o tym samym lemacie (różniących się parametrami morfosyntaktycznymi) nazwiemy paradygmatem odmiany.

O lematach

Konwencja

Dla każdej części mowy umawiamy się, że pewną jej formę nazwiemy lematem (formą bazową).

- Dla rzeczownika jest to mianownik liczby pojedynczej, dla kuleczkami lematem jest kuleczka.
- Dla czasownika lematem jest bezokolicznik (powiedziałbyś – powiedzieć).
- Dla przymiotnika lematem jest mianownik rodzaju męskiego liczby pojedynczej (najpiękniejszej – najpiękniejszy)
- Dla nieodmiennych wyrazów one same są swoimi lematami.

Wszyscy posługujemy się lematami: na przykład wyszukując informacji w Internecie.

O lematach (2)

Oczywiście można wybrać inaczej. W szczególności:

- Moglibyśmy się umówić, żeby dla pociągającego lematem był na przykład pociągać,
- a dla najpiękniejszej – piękny (albo najpiękniejsza)

O lematach (3)

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi
- mam
- barki
- **tonie**
- winie

O lematach (3)

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi – musieć, muszy (mający związek z muchą)
- mam – mama, mieć, mamić
- barki – barka, bark, barek
- **tonie** – toń, tonąć, tona, ton
- winie – wina, wino

Znaczenie lematów w wyszukiwaniu

- Wyszukiwarki zwyczajowo utożsamiają słowa o tym samym lemacie (być może nieco preferując brzmienie dosłowne)
- W przypadku słów wielolematowych może to sprawić problem: jak znaleźć słowo 'barek' w Wikipedii (rozumiany jako szafka na alkohol)
- Pewnym problemem jest również analiza zdań ze słowami wielolematowymi (zwłaszcza popularnymi, takimi jak 'je', 'lub', 'albo')

Lematy i modelowanie języka

Uwaga

Patrzenie na następstwa lematów może pomóc modelowi językowemu, bo występująca w korpusie:

wesoła wiewiórka

uprawdopodobnia frazy:

wesołych wiewiórek, *wesołymi wiewiórkami*, *wesoła wiewiórko*...

Takie wielowariantowe połączenia są również często silnymi kolokacjami!

Jak policzyć statystyki dla lematów?

Rozwiązanie 1

Ujednoznacznić tekst, potem po prostu policzyć.
Wada: wszystko potem będziemy musieli ujednoznaczniać, co może być trochę niewygodne.

Rozwiązanie 2

Wprowadzić pojęcie Jednoznacznego Prawie Lematu (nazwa nasza). Jak?

JPL dla słowa to zbiór lematów tego słowa.

JPL: przykłady i konsekwencje

- W oczywisty sposób nic nie tracimy (najwyżej JPL będzie dziwną nazwą słowa)
- Ale w praktyce zdecydowana większość (98%) słów ma jednoznaczne bazy, więc JPL jest singletonem zawierającym po prostu lemat słowa.
- Popatrzmy na taki zlematyzowany tekst:

  Lingwistyka stosowana – dziedzina nauki zajmująca się rozpoznawaniem i badaniem zagadnień związanych z językiem oraz rozwiązywaniem występujących w praktyce problemów językowych. Pokrewne jej działy to: dydaktyka, językoznawstwo, psychologia, antropologia oraz socjologia.

JPL: wynik lematyzacji

lingwistyka stosowany – dziedzina nauka ***zajmująco_zajmujący*** *się rozpoznawanie i badanie zagadnienie związany z język oraz rozwiązywanie występujący w praktyka problem językowy . pokrewny on* **dziać_dział** *ten :* **dydaktyk_dydaktyka** *, językoznawstwo , psychologia , antropologia oraz socjologia*

- Tekst jest generalnie czytelny (co dobrze rokuje dla word2vec-a)
- Widać błąd w słowniku – zajmująco nie odmienia się jak gorąco (gorąca, gorącem)

Gender dla dzieci

- W szkole mówiono o rodzajach: męskim, żeńskim i nijakim (m, f, n)
- A w liczbie mnogiej z kolei o rodzaju męskoosobowym i niemęskoosobowym
- To niekomfortowa sytuacja, bo nie możemy powiedzieć: rzeczownik ma rodzaj G.

Gender dla dorosłych (1)

- Zakładamy, że rodzaj jest właściwością pewnych lematów i nie zależy on od innych parametrów słowa.
- Rodzaj powinien opisywać zachowanie i umożliwiać decyzję, jaki przymiotnik może połączyć się z jakim rzeczownikiem (przy założeniu, że mają mieć wspólne liczbę, przypadek i rodzaj).
- Będziemy używać już prawdziwych oznaczeń, które spotykamy w korpusach.

Gender dla dorosłych (2)

- Potrzebujemy trzech rodzajów męskich: osobowego (m1), zwierzęcego (m2), rzeczowego (m3). Porównajmy
  - Widzę pięknego strażaka. Widzę pięknego rumaka. Widzę piękny hamak.
  - Biegną piękni strażacy. Biegną piękne rumaki. Biegną piękne hamaki.
- Mniej niż 3 rodzaje nie pozwolą poprawnie opisać powyższych konstrukcji.
- Niektórzy wprowadzają również 2 rodzaje nijakie, ze względu na odmienne połączenie z liczebnikami: dwa okna (n2), dwoje kurcząt (n1)
- Czasami używa się również rodzajów p1,p2,p3 dla rzeczowników, które występują tylko w liczbie mnogiej, takich jak spodnie, dane, czy wujostwo.

Mały bonus

Mamy dwie ciekawe klasy zdeterminowane gramatycznie (a jednocześnie trochę semantyczne):

Rodzaj m1

idiota alopata chłopaczek filipińczyk szalbierz Włodzimierz faworyt farbiarz mahatma obwieś fotoamator Rumun autysta emeryt społecznik kobieciarz cinkciarz sejsmolog stalinowiec parszywiec agitator nowochrzczeniec jazzman magister ośmiolatek Petrarka kacyk muzykoterapeuta

Rodzaj m2

wróbelek gawron karlik krabołów dzieciak bobak bażancik omułek koprofag marabut żubrobizon bielik kangurek bułanek maluszek raczek inochodziec

Tagi dla języka polskiego

- Różne zbiory tagów dla języka polskiego są dość podobne.
- Różnice dotyczą:
  - rodzajów (panuje zgoda co do tego, że konieczne są rodzaje f, n, m1, m2, m3)
  - traktowania takich słów jak *przeczytalibyście* jako jednego segmentu, lub trzech: przeczytali-by-ście
  - wyrazów takich jak pierwszy (liczebnik porządkowy) i były (specjalny rodzaj imiesłowu przymiotnikowego) – mogą być one osobnymi częściami mowy, lub przymiotnikami
  - takich rzeczowników jak czytanie i pisanie (gerundium). Niekiedy uznawane są one za odmienną część mowy.

Tagset

Tagset

Tagsetem nazwiemy zbiór znaczników służących do opisu gramatycznych właściwości wyrazów. Tag składa się z:

- identyfikatora części mowy,
- parametrów tej części, podanych we właściwej dla danej części mowy kolejności.
- Oddzielamy je zwyczajowo dwukropkami

Części mowy

Podstawowe części mowy to: rzeczownik (subst), czasownik (verb), przymiotnik (adj), przysłówek (adv). Ważne są też imiesłowy (pact-czytający, ppas-czytany, pcon-czytając, pant-przeczytawszy).

Kilka nietypowych części mowy i tagów

- Wprowadzona została specjalna część mowy (kublik), zawierająca takie „specjalne” słowa jak: się, nie, ba, och ...
- Jest specjalna, nieodmienna część mowy: rzeczownik deprecjatywny (np. profesory, doktory). Można patrzeć na nią jak na uzwierzęcenie (m2) wyrazów w rodzaju (m1). Oznaczamy tagiem depr.
- Dla słów *winien* i *powienien* utworzono osobną część mowy (odmieniającą się przez liczby i rodzaje).
- przyimki mają specjalny parametr, wokaliczność, pozwalający odróżnić słowa *nad* oraz *nade*.

Słownik morfosyntaktyczny

- Podstawową funcjonalnością jest pamiętanie słów wraz z opisem gramatycznym.
- Można to zrobić inteligentniej, za pomocą drzewa trie, albo word-dagu, ale nie jest to na tyle duża strukrtura, żeby nie dało się jej wczytać w pamięci.
- Są dwa (co najmniej) powszechnie używane, darmowe, słowniki:
  - Słownik Gramatyczny Języka Polskiego (sgjp.pl/morfeusz)
  - Morfologik (morfologik.blogspot.com)
  - Polimorf (połączenie wyżej wymienionych)

Jak wyglądał plik morfologik.txt

```
dziewczyn    dziewczyna subst:pl:gen:f
dziewczynie  dziewczyna subst:sg:dat.loc:f
dziewczyno   dziewczyna subst:sg:voc:f
dziewczynom  dziewczyna subst:pl:dat:f
dziewczyny   dziewczyna subst:sg:gen:f
dziewczyną   dziewczyna subst:sg:inst:f
dziewczynę   dziewczyna subst:sg:acc:f
piękna piękno subst:pl:nom.acc:n+subst:sg:gen:n
piękna piękny adj:sg:nom:f:pneg
niearcypiękna arcypiękny adj:sg:nom:f:neg
```

Pełność słowników

- Słownik SGJP chwali się 320 tys haseł (oczywiście, ze względu na odmianę, słów w słowniku jest dużo więcej)
- Przeprowadzimy prosty test: wypisujemy dużo słów, raczej niezbyt popularnych i sprawdzamy, ile z nich znajduje się w słowniku.

Pełność SGJP

Na czerwono słowa, których nie ma w słowniku. Na zielono, te co są, choć to trochę zaskakuje :)

ściema zarąbisty full wypas pijarowy teczuszka paprotka szybkościowość pompeczka brzuszek wymyk hydroksyzyna przeciwwstrząsowy filatelistyczny twierdzenie pitagorasa przyprostokątna przeciwprostokątna perłopławy ośmiorniczki grillowanie przysmażony odsmażany zagęścić wyguglować drzewiasty haszowanie haszujący horror horrorystyczny menuet tańcować przetańcować przekabacić przyfasolić przyprawianie grupoid monoid sekretareczka fantasmagoria krasnolud elf elfi po krasnoludzku wiedźmin gizarma morgenstern miecz obosieczny topór oburęczny półpancerz misiurka

Pełność SGJP

Na czerwono słowa, których nie ma w słowniku. Na zielono, te co są, choć to trochę zaskakuje :)

ściema zarąbisty full wypas pijarowy teczuszka paprotka szybkościowość pompeczka brzuszek wymyk hydroksyzyna przeciwwstrząsowy filatelistyczny twierdzenie pitagorasa przyprostokątna przeciwprostokątna perłopławy ośmiorniczki grillowanie przysmażony odsmażany zagęścić wyguglować drzewiasty haszowanie haszujący horror horrorystyczny menuet tańcować przetańcować przekabacić przyfasolić przyprawianie grupoid monoid sekretareczka fantasmagoria krasnolud elf elfi po krasnoludzku wiedźmin gizarma morgenstern miecz obosieczny topór oburęczny półpancerz misiurka

Pełność Polimorfa

Na czerwono słowa, których nie ma w słowniku. Na zielono, te co są, choć to trochę zaskakuje :)

ściema zarąbisty full wypas pijarowy teczuszka paprotka szybkościowość pompeczka brzuszek wymyk hydroksyzyna przeciwwstrząsowy filatelistyczny twierdzenie pitagorasa przyprostokątna przeciwprostokątna perłopławy ośmiorniczki grillowanie przysmażony odsmażany zagęścić wyguglować drzewiasty haszowanie haszujący horror horrorystyczny menuet tańcować przetańcować przekabacić przyfasolić przyprawianie grupoid monoid sekretareczka fantasmagoria krasnolud elf elfi po krasnoludzku wiedźmin gizarma morgenstern miecz obosieczny topór oburęczny półpancerz misiurka

Czy więcej znaczy lepiej?

Przeprowadźmy analizę morfosyntaktyczną zapytania:

Czy zawsze lepiej używać polimorfa?

Analizator uruchamiamy pisząc: morfeusz_analyzer -d polimorf

Uwaga

Morfeusz ma również interfejs programistyczny, również w Pythonie. Więcej na pracowni.

Wzbogacanie słownika

- Słownik nigdy nie będzie pełny, a w zadaniu oczekujemy tego, żeby on pełny był.
- Można wzbogacać słownik o reguły, na przykład:
  - tag(-ając) =
  - tag(-awszy) =
  - tag(-ywa) =
  - tag(-ający) =

Wzbogacanie słownika)

- Słownik nigdy nie będzie pełny, a w zadaniu oczekujemy tego, żeby on pełny był.
- Można wzbogacać słownik o reguły, na przykład:
  - tag(-ając) = tag(zając) + tag(czytając)
  - tag(-awszy) = tag(przeczytawszy)
  - tag(-ywa) = tag(recydywa) + tag(wydobywa)
  - tag(-ający) = tag(czytający)