Osadzenia bezkonteksowe i kontekstowe

Paweł Rychlikowski

Instytut Informatyki UWr

20 listopada 2024

Cele na dziś (częściowo podobne, co tydzień temu)

Nasz cel (nie Norwida)

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Takie przypisanie nazwiemy **osadzaniem** słów/tokenów

Najważniejsze dla nas będą **osadzenia kontekstowe** i zrozumienie, co z nich wynika dla **Przetwarzania języka naturalnego (NLP)**

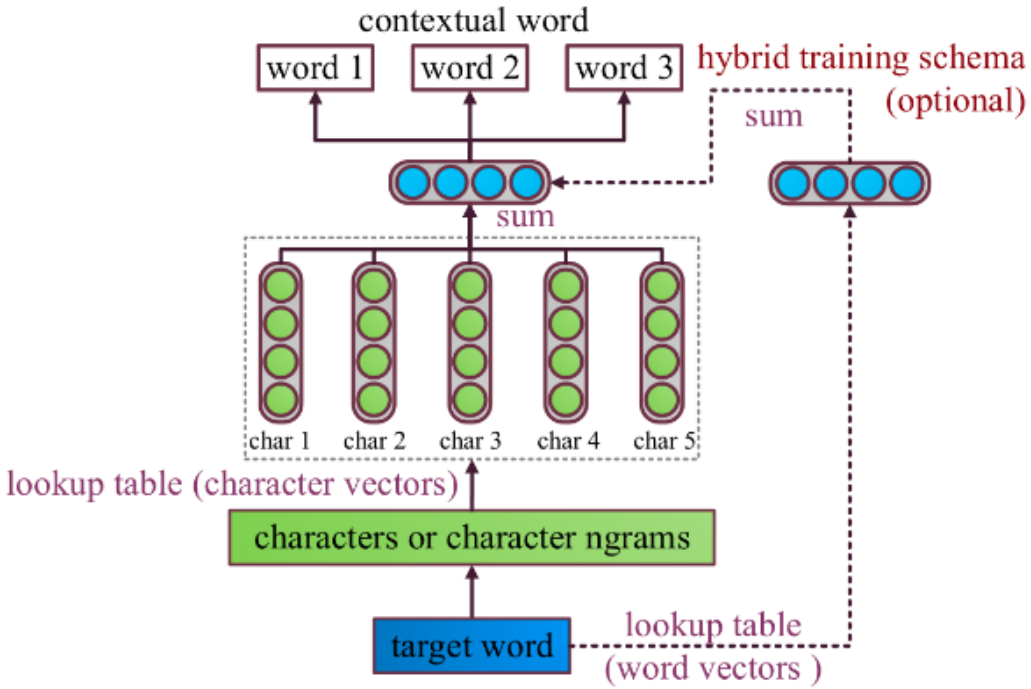
Zanurzenia części słów

Problem

Nawet największy Word2Vec nie daje gwarancji, że obejmie wszystkie słowa (słowa fachowe, literówki, nazwy własne, słowa z innych języków, ...)

- Można zastosować word2vec dla stokenizowanego tekstu (np. za pomocą BPE)
- Zanurzenie słowa = suma zanurzeń tokenów (może z jakimiś wagami)
- (za chwilę zobaczymy, jak mogłoby to działać)

Zanurzenia części słów (fasttext)



Żródło: https://vecto.space/projects/subword

Uwaga

Fasttext do pewnego stopnia działa jako narzędzie do usuwania literówek! (czy wiadomo dlaczego?)

Reprezentacja dokumentu

- Potrzebujemy sposobu na reprezentację dokumentu: choćby w sytuacji, w której mamy osadzenia tokenów, a chcemy mieć reprezentację słowa.
- Oczywisty pomysł – suma wektorów
- Ale warto trochę się cofnąć, i rozważyć nie tylko gęste, ale i rzadkie reprezentacje dokumentów.

Bag-of-words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Okazuje się, że zapominanie o kolejności słów w wielu sytuacjach jest ok (przykłady na kolejnym slajdzie).

Bag-of-words. Kiedy działa?

- **Działa:**
  - Klasyfikacja dziedziny, segregacja newsów, czy artykułów naukowych
  - Ustalanie języka tekstu (pomógłby tokenizator wielojęzyczny)
  - Do pewnego stopia: ustalanie prostoty tekstu (tylko słowa do B2)
  - Ogólnie: proste zadania na całym tekście
- **Nie działa:**
  - Wydźwięk: on nie jest nie fajny (sic!) vs nie, on nie jest fajny
    - ★ Zagadka: dlaczego w klasyfikacji wydźwięku recenzji telefonów użyteczne były słowa: telefon i telefonu.
  - Fake news detection (i inne zadania nietrywialnej klasyfikacji)
  - Ocena jakości tekstu (styl, sensowność, spójność argumentacji, ...)
  - i wiele innych

Referencyjna metoda to Naive Bayes Classifier (ew. regresja liniowa).

Istotność słów

- Nie wszystkie słowa są tak samo ważne: zbadano vs hemoglobina
- Podstawowa intuicja: **nieistotne są słowa, występujące „prawie wszędzie”**
- Można spróbować zrobić ich listę

Stopwords Wikipedia

a, aby, ach, acz, aczkolwiek, aj, albo, ale, ależ, ani, aż, bardziej, bardzo, bo, bowiem, by, byli, bynajmniej, być, był, była, było, były, będzie, będą, cali, cała, cały, ci, cię, ciebie, co, cokolwiek, coś, czasami, czasem, czemu, czy, czyli, daleko, dla, dlaczego, dlatego, do, dobrze, dokąd, dość, dużo, dwa, dwaj, dwie, dwoje, dziś, dzisiaj, gdy, gdyby, gdyż, gdzie, gdziekolwiek, gdzieś, i, ich, ile, im, inna, inne, inny, innych, iż, ja, ją, jak, jaka, jakaś, jakby, jaki, jakichś, jakie, jakiś, jakiż, jakkolwiek, jako, jakoś, je, jeden, jedna, jedno, jednak, jednakże, jego, jej, jemu, jest, jestem, jeszcze, jeśli, jeżeli, już, ją, każdy, kiedy, kilka, kimś, kto, ktokolwiek, ktoś, która, które, którego, której, który, których, którym, którzy, ku, lat, lecz, lub, ma, mają, mało, mam, mi, mimo, między, mną, mnie, mogą, moi, moim, moja, moje, może, możliwe, można, mój, mu, musi, my, na, nad, nam, nami, nas, nasi, nasz, nasza, nasze, naszego, naszych, natomiast, natychmiast, nawet, nią, nic, nich, nie, niech, niego, niej, niemu, nigdy, nim, nimi, niż, no, o, obok, od, około, on, ona, one, oni, ono, oraz, oto, owszem, pan, pana, pani, po, pod, podczas, pomimo, ponad, ponieważ, powinien, powinna, powinni, powinno, poza, prawie, przecież, przed, przede, przedtem, przez, przy, roku, również, sama, są, się, skąd, sobie, sobą, sposób, swoje, ta, tak, taka, taki, takie, także, tam, te, tego, tej, temu, ten, teraz, też, to, tobą, tobie, toteż, trzeba, tu, tutaj, twoi, twoim, twoja, twoje, twym, twój, ty, tych, tylko, tym, u, w, wam, wami, was, wasz, wasza, wasze, we, według, wiele, wielu, więc, więcej, wszyscy, wszystkich, wszystkie, wszystkim, wszystko, wtedy, wy, właśnie, z, za, zapewne, zawsze, ze, zł, znowu, znów, został, żaden, żadna, żadne, żadnych, że, żeby

Istotność słów

- Można to jakoś subtelniej stopniować liczbowo (jak?)

Definicja

Inverted Document Frequency (IDF) wyraża się wzorem:

$$\mathrm{IDF}(w) = \log(\frac{N}{\mathrm{cnt}(w)})$$

gdzie N jest liczba zdań (dokumentów, akapitów, słów) w korpusie, natomiast $\mathrm{cnt}(w)$ jest liczbą zdań (etc.), zawierających w

Rzadka i gęsta wektorowa reprezentacja dokumentu

Reprezentacja 1

rzadki wektor, na niezerowych pozycjach liczby wystąpień słowa pomnożene przez jego IDF (**TF-IDF**)

- Dla tak określonych dokumentów oczywiście można obliczać cosinus, jest on uciągle powszechnie żywaną miarą podobieństwa)
- Popularny wariant: BM-25 (ćwiczenia)

Reprezentacja 2

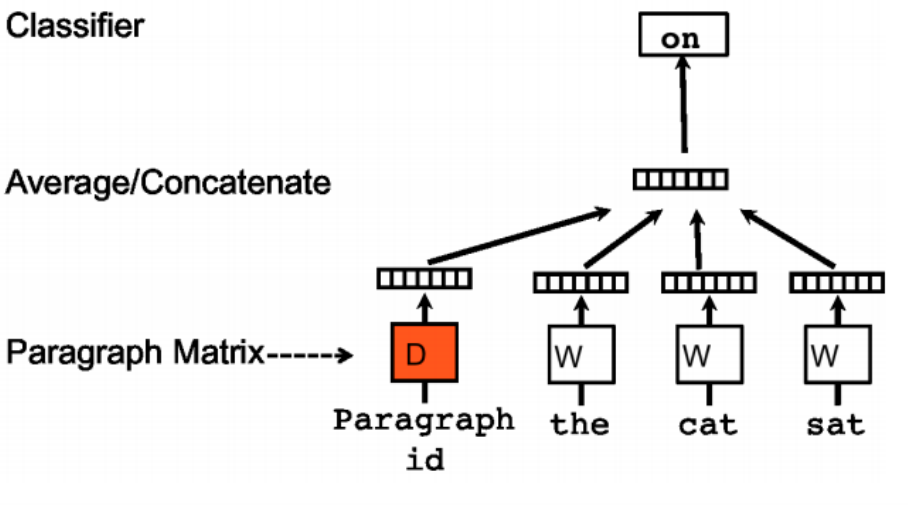
Bag-of-Vectors (BoV)

- Po prostu suma wektorów słów (ewentualnie ważonych przez **IDF**)
- Zwróćmy uwagę, że TF-IDF też jest sumą wektorów (one-hot)

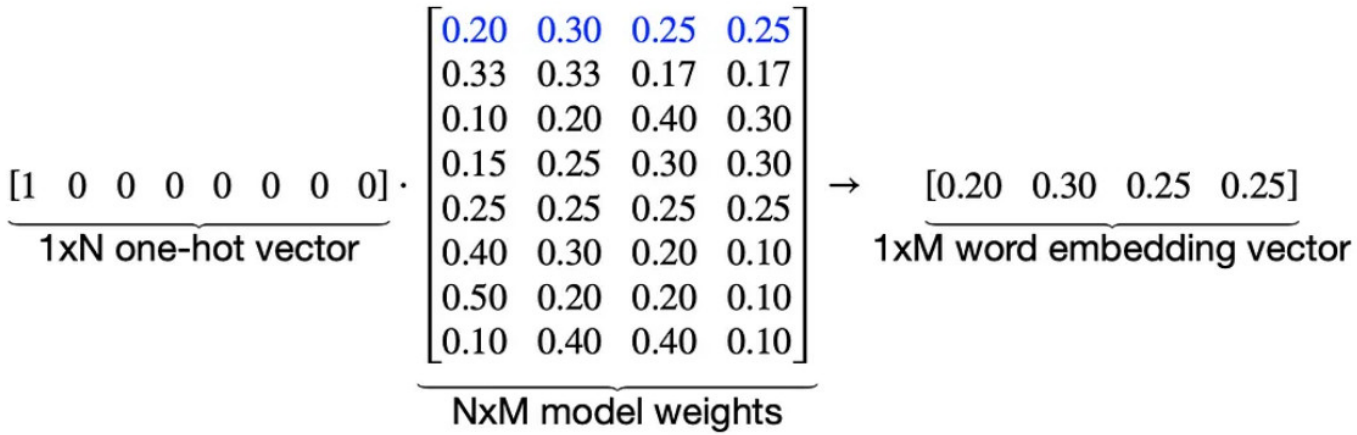
Łączne reprezentacje słow i dokumnentów: doc2vec

- Pierwsze przybliżenie: w tekście co jakiś czas dać identyfikator dokumentu
- Drugie przybliżenie: doc2vec (od twórców word2vec)

Praca: Distributed Representations of Sentences and Documents (Quoc Le, Tomas Mikolov)



Osadzenia jako część innych sieci neuronowych



- Osadzenia są związane z kodowaniem one-hot (rysunek powyżej)
- Oczywiście w praktyce tego mnożenia się nie wykonuje, tylko odczytuje wektor z tablicy

Uwaga

Warstwa osadzeń występuje w zasadzie w każdej sieci neuronowej robiącej coś z tekstami (lub ciągami elementów ze skończonego zbioru)

Popatrzmy na notatnik embeddings.ipynb

Bezkontekstowe vs kontekstowe osadzenia

Bezkontekstowe (word2vec)

- **PLUS**: jakaś część znaczenia słowa jest niezależna od kontekstu (bo inaczej nie moglibyśmy się porozumiewać
- **MINUS**: ale word2vec **zawsze** pomija kontekst!

Słowa wieloznaczne lub uzyskujące znaczenie

- ta potrawa, ta autorka, nasz bohater, ...
- Słowa wieloznaczne: zamek, żabka, stan, dół, szczyt
- Polisemia: bank (budynek, instytucja, firma)
- Te zjawiska zachodzą w różnych językach (bank, lead, bass, content, ...)

Ponadto, jak widzieliśmy w demonstracji, niektóre tokeny są „zbyt krótkie", żeby nieść ze sobą znaczenie

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

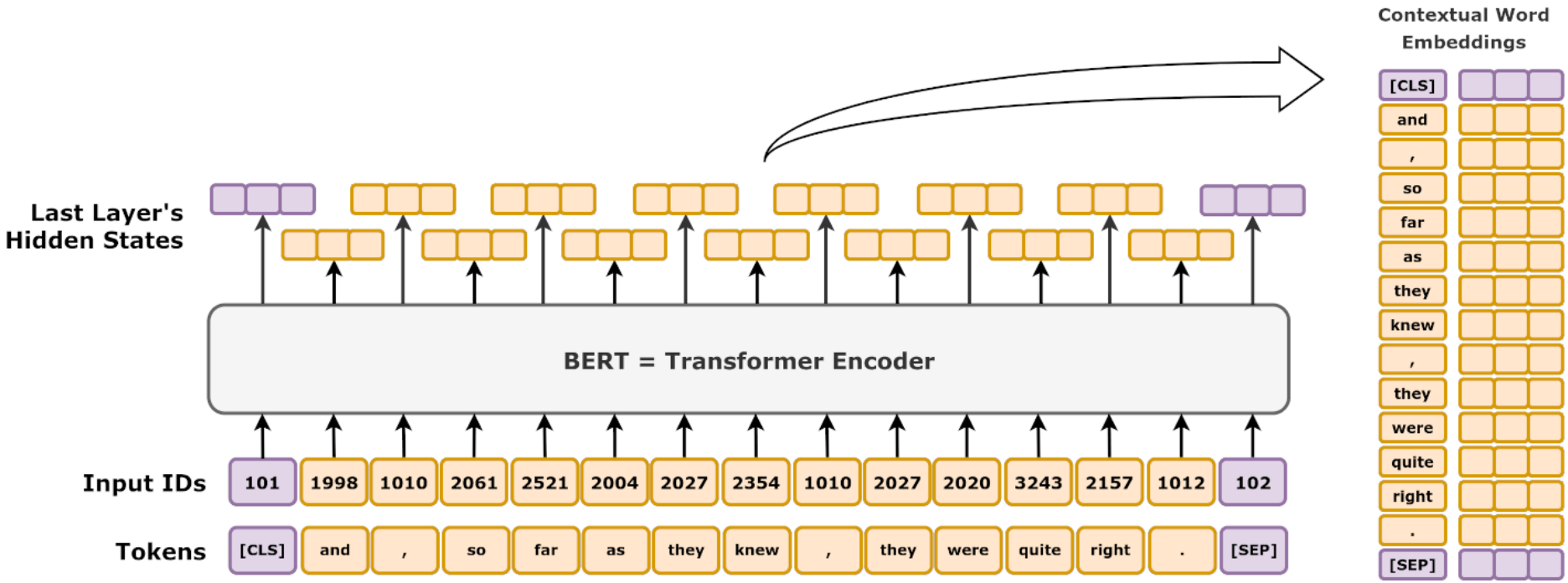
Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language repre-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that

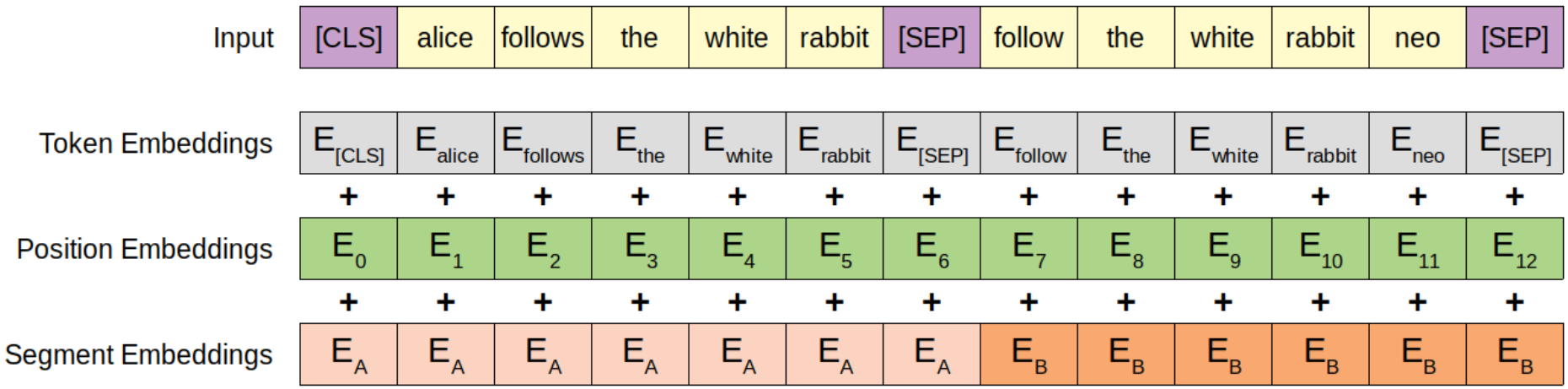
- Najbardziej udany koder transformerowy!
- Praca, która zmieniła definitywnie NLP!

Ogólny schemat BERT-a



Kodowanie wejścia BERT-a

- Sieci transformer przetwarzają wejście równolegle, traktując każdy wektor tak samo.
- Oczywiście kolejność słów ma znaczenie, trzeba ją jakoś zakodować



Rzut oka na kodowania pozycji w Papudze

Popatrzmy ponownie na notatnik embeddings.ipynb

BERT jako ekstraktor cech

- Możemy potraktować wektor wynik obliczony przez BERT-a, jako reprezentację zdania.
  - Albo token dla pozycji |<CLS>|
  - Albo średnią wektorów dla tokenów

BERT po polsku

- Istnieje kilka powszechnie dostępnych modeli typu BERT dla języka polskiego (np. HerBERT, PolBERT)
- Zobaczymy jak Herbert działa w najprostszym scenariuszu
  - Przerwa na demonstracje (`herbert.ipynb`)
- Na liście 3 zbadamy bardziej złożone warianty:
  - Różne inne algorytmy ML (+regularyzacja)
  - Augmentacja danych
  - Łączenie Herberta z Papugą