Szemantikus reprezentáció magyar nyelv esetén

Kántor Attila, Grad-Gyenge László

2020. június 23.

A projekt az Európai Unió támogatásával,
az Európai Szociális Alap társfinanszírozásával
valósult meg
(EFOP-3.6.3-VEKOP-16-2017-00002).

Motiváció

- Természetesnyelv-feldolgozás részterülete
- Nyelvi elemek numerikus ábrázolása (szólista $\rightarrow$ vektortér)
- Modern módszerek meghatározó elemei:
  - Alapjául szolgáló neurális háló (általában rekurrens vagy rekurzív)
  - Tanítási feladatok és adathalmazok
- Jó reprezentáció esetén közel kerülnek az azonos jelentésű vektorok egymáshoz

Motiváció

- Az így kapott vektorok számos módon felhasználhatók (osztályozás, keresőmotorok, chatbot)
- Léteznek többnyelvű megoldások, de a többség nyelvfüggő
- A modellek tanításához nagy mennyiségű adat szükséges
- Ember által címkézett adat $\rightarrow$ pontosabb eredmény
- Kis és közepes nyelvek (magyar) problémája: limitált eszköztár (csak felügyelet nélküli tanítás)
- Meglévő magyar nyelvű módszerek
  - Csak szavak szintjén $\rightarrow$ kevésbé pontos
  - Nincs lehetőség egyetlen modell segítségével nagyobb nyelvi elemek feldolgozására

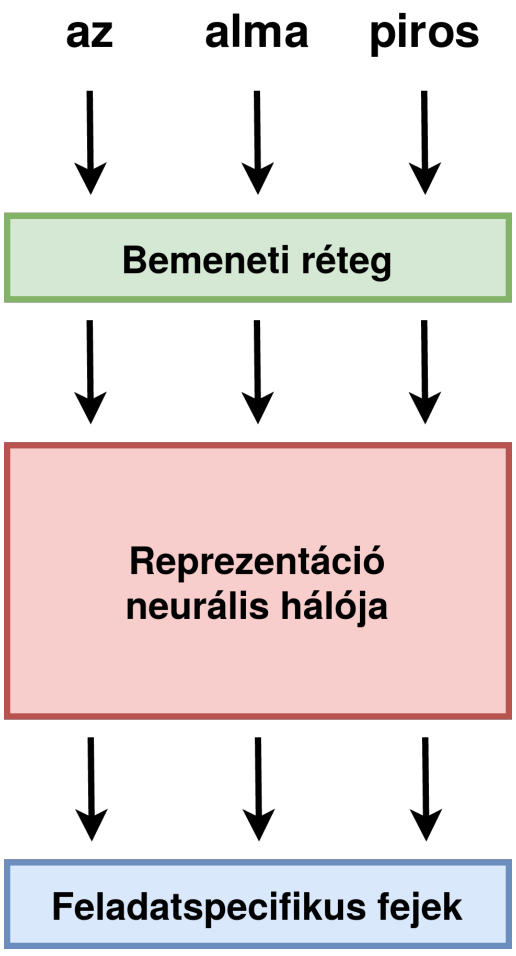
Előzmények

- Szóbeágyazási módszerek (szó $\rightarrow$ vektor)
  - Word2Vec, GloVe
  - Lokális és/vagy globális statisztikák
- InferSent: Ember által annotált adat $\rightarrow$ jobb teljesítmény
- BERT: Autoannotált adattal is State-of-the-art eredmények
- USE: Kevés adat esetén jó megoldás lehet a transfer learning módszere

Felhasznált adathalmazok

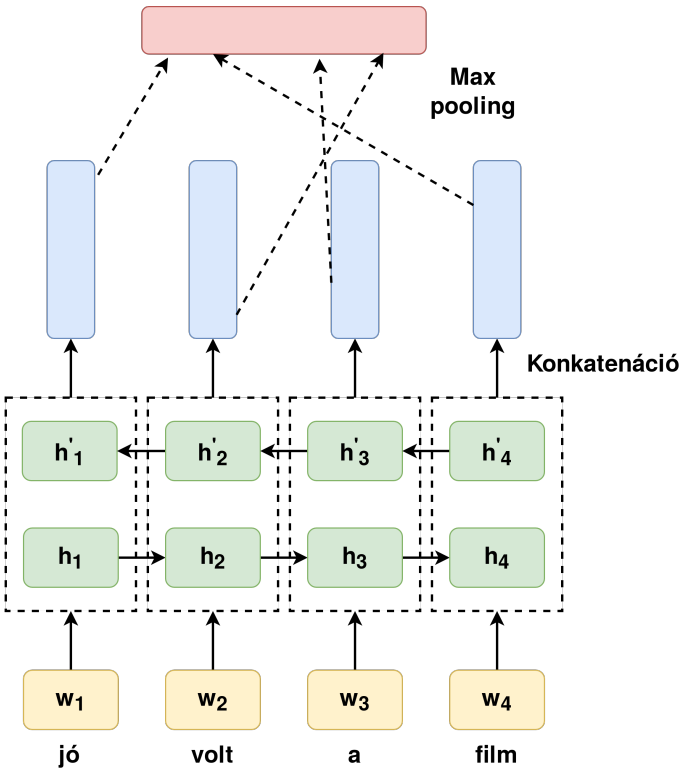
- A magyar nyelven elérhető tanítóhalmazok és források száma igen csekély (publikus korpuszok, web)
- OSCAR (40GB, csak a felét használtam) - publikus
  - A Common Crawl tisztított és klasszifikált változata
  - A mondatok nem sorrendtartóak
- Hungarian Webcorpus (6,5 GB) - publikus
  - Eredendően magyar nyelvű adathalmaz
  - Kb. 1,2 millió magyar weboldal
- Vásárlói vélemények (95 000 bejegyzés)
  - Saját halmaz (Árukereső), vélemény - csillagok száma párosok
  - 1-2 csillag $\rightarrow -$, 4-5 csillag $\rightarrow +$
  - Tisztítva, egyensúlyozva

Az előtanítás - Architektúra



- Generált bemenet a Hungarian Webcorpus-ból
- Bemeneti tokenszekvencia Word2Vec vektorai (OSCAR - 645 000 méretű szótár)

Az előtanítás



- BiLSTM max pooling (InferSent)
- Kétirányú olvasás
- Feladatok (BERT):
  - **Maszkolás**: takarjuk le a szavak 15%-át, a háló feladata kitalálni az eredeti tokeneket
  - **Következő mondat**: A és B mondatok, vajon B szekvencia A után következik az eredeti dokumentumban?

Az előtanítás

- Ismeretlen szavak problémája $\rightarrow$ saját dimenzió
- Normált bemenet $\rightarrow$ jobb teljesítmény
- LSTM rejtett méret növelése

Rejtett méret	**Maszkolás**	**Következő mondat**
1024	16.61%	60.49%
4096	**17.82%**	**94.15%**

táblázat: Az előtanítás teszt pontossága

A módszer kiértékelése

- A jó kiértékelési feladat jellemzői:
  - Kellően nehéz, jól láthatóak a különbségek
  - Jól interpretálható végeredmény
- Vásárlói vélemények bináris klasszifikációja érzelmi tartalom alapján megfelelő
- Viszonyítási alap: Word2Vec vektorok átlaga
- Mérés menete: mondatvektorok generálása $\rightarrow$ osztályozó algoritmus $\rightarrow$ eredmény kimérése teszthalmazon

A módszer kiértékelése

Modell / Osztályozó	Linear SVM	XGBoost	Random Forest
w2v_sm	85,45%	82,18%	83,64%
w2v_sm_norm	85,57%	82,71%	**84,18%**
w2v_lg	85,28%	82,87%	83,77%
w2v_lg_norm	85,59%	83,15%	83,85%
lstm_1024	81,53%	80,82%	80,84%
lstm_1024_norm	85,37%	83,48%	83,41%
lstm_4096_norm	**87.16%**	**83,71%**	83,90%

táblázat: A modellek klasszifikációs pontossága a vélemények adathalmazon végzett bináris osztályozási feladat esetén. A w2v_sm az oscar_sm, a w2v_lg az oscar halmazon tanított modelleket, a norm posztfix a normált bemenetet jelöli. A számok a modellek nevében a reprezentációs vektor méretére utalnak.

További fejlesztések - a dolgozat nem tartalmazza

- Ismeretlen tokenek 0 súlyozása a maszkolási feladat esetén $\rightarrow$ teljesítmény romlott
- lstm_4096_norm modell finomhangolása a vélemények adathalmazon $\rightarrow$ jelentős javulás (89,67 % teszt pontosság)

Összefoglaló

A dolgozat tartalma:

- Angol nyelvű módszerek és magyar nyelvű korpuszok vizsgálata
- Egy előre tanított magyar nyelvű kétirányú reprezentációs modell létrehozása mondatokra és paragrafusokra
- Mérési adathalmaz magyar nyelvű reprezentációs módszerekhez
- A bemutatott algoritmus az adott feladat esetén túlteljesítette a Word2Vec-et

Köszönöm a figyelmet!

SZÉCHENYI 2020

MAGYARORSZÁG KORMÁNYA

Európai Unió
Európai Szociális Alap

BEFEKTETÉS A JÖVŐBE