Entitás kinyerés magyar nyelvű szövegekből kétirányú LSTM-mel

Scheier Balázs – FAKK40, Mészáros Bálint – HY90XY

Budapesti Műszaki és Gazdaságtudományi Egyetem

Entitás felismerés bevezető

- NLP feladat

Entitás felismerés bevezető

- NLP feladat
- Célja, hogy felismerjük a szövegben szereplő előre maghatározott kategóriába tartozó entitásokat.

Entitás felismerés bevezető

- NLP feladat
- Célja, hogy felismerjük a szövegben szereplő előre maghatározott kategóriába tartozó entitásokat.
- Fontos alapfeladat a szövegfeldolgozásban.

Az entitásokról

A következő entitásokkal foglalkozunk:

- Helyszín (location)
- Személy (person)
- Szervezet (organization)
- Vegyes (miscellaneous)

Az entitásokról

A következő entitásokkal foglalkozunk:

- Helyszín (location)
- Személy (person)
- Szervezet (organization)
- Vegyes (miscellaneous)

Jelölések:

- B-LOC, I-LOC
- B-PER, I-PER
- B-ORG, I-ORG
- B-MISC, I-MISC
- 0

Az entitásokról

A következő entitásokkal foglalkozunk:

- Helyszín (location)
- Személy (person)
- Szervezet (organization)
- Vegyes (miscellaneous)

Jelölések:

- B-LOC, I-LOC
- B-PER, I-PER
- B-ORG, I-ORG
- B-MISC, I-MISC
- 0

Budapest egy szép város. – B-LOC
A budapesti kirándulás nagyon tetszett. – Nem entitás

Adatbázis

- hunNERwiki

```
A           text    0       ART     a       O
céljuk      text    0       NOUN<POSS<PLUR>>        cél     O
,           text    0       PUNCT   ,       O
hogy        text    0       CONJ    hogy    O
biztosítsák     text    0       VERB<SUBJUNC-IMP><PLUR><DEF>    biztosít        O
,           text    0       PUNCT   ,       O
hogy        text    0       CONJ    hogy    O
a           text    0       ART     a       O
korábbi     text    0       ADJ     korai   O
szerzők     text    0       NOUN<PLUR>      szerző  O
kilétét     text    0       NOUN<POSS><CAS<ACC>>    kilét   O
```

Az adat struktúrája

Architektúra

```
'B-LOC' : 0,
'B-MISC' : 1,
'B-ORG' : 2,
'B-PER' : 3,
'I-LOC' : 4,
'I-MISC' : 5,
'I-ORG' : 6,
'I-PER' : 7,
'O' : 8,
'PAD' : 9,
'BOS' : 10,
'EOS' : 11
```

(a) A szótárunk

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 28)]              0

 embedding (Embedding)       (None, 28, 64)            1262464

 bidirectional (Bidirectiona  (None, 28, 512)          657408
 l)

 bidirectional_1 (Bidirectio  (None, 512)              1574912
 nal)

 dense (Dense)               (None, 12)                6156

=================================================================
Total params: 3,500,940
Trainable params: 3,500,940
Non-trainable params: 0
_________________________________________________________________
```

(b) A neurális háló architektúrája

Tanítás

- Adatok: tanító-teszt-validációs: 0.6-0.2-0.2

Tanítás

- Adatok: tanító-teszt-validációs: 0.6-0.2-0.2
- Optimalizáció accuracy-ra

Tanítás

- Adatok: tanító-teszt-validációs: 0.6-0.2-0.2
- Optimalizáció accuracy-ra
- Early stopping 5 epoch után

Hiperparaméter optimalizálás

- Rejtett rétegek mérete: 64, 128, 256, 512

Hiperparaméter optimalizálás

- Rejtett rétegek mérete: 64, 128, 256, 512
- Optimalizációs eljárások: RMSProp, Adam, SGD

Hiperparaméter optimalizálás

- Rejtett rétegek mérete: 64, 128, 256, 512
- Optimalizációs eljárások: RMSProp, Adam, SGD
- Batch méret: 32, 64, 128, 256

Hiperparaméter optimalizálás

Embed-ding réteg mérete	Első LSTM réteg mérete	Második LSTM réteg mérete	Optimalizációs algoritmus	Batch méret	Legjobb validációs accuracy
64	256	256	Adam	256	0.9905
512	256	512	RMSprop	256	0.9883
512	512	64	Adam	64	0.9881
512	512	512	RMSprop	256	0.9876
128	128	128	Adam	128	0.9875

Az 5 legjobb eredmény

Eredmények

A legjobb modellel (64, 256, 256, Adam, 256)

Eredmények

A legjobb modellel (64, 256, 256, Adam, 256)

- Tanító accuracy: 0.9973

Eredmények

A legjobb modellel (64, 256, 256, Adam, 256)

- Tanító accuracy: 0.9973
- Validációs accuracy: 0.9861

Eredmények

A legjobb modellel (64, 256, 256, Adam, 256)

- Tanító accuracy: 0.9973
- Validációs accuracy: 0.9861
- Teszt accuracy: 0.9874

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-PER', 'I-PER', 'O', 'O'

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-PER', 'I-PER', 'O', 'O'
  - A predikció:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O'

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-PER', 'I-PER', 'O', 'O'
  - A predikció:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O'
- Öt ok, amiért Macron Budapestre látogat.

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-PER', 'I-PER', 'O', 'O'
  - A predikció:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O'
- Öt ok, amiért Macron Budapestre látogat.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'B-PER', 'B-LOC', 'O'

Demo

- A tudomány szerint is működik az alvásmódszer, amit Salvador Dalí is használt.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-PER', 'I-PER', 'O', 'O'
  - A predikció:
    'BOS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O'
- Öt ok, amiért Macron Budapestre látogat.
  - Az eredeti címkék:
    'BOS', 'O', 'O', 'O', 'B-PER', 'B-LOC', 'O'
  - A predikció:
    'BOS', 'O', 'O', 'O', 'B-LOC', 'B-LOC', 'O'

Összefoglalás

- Téma: Entitás kinyerés magyar nyelvű szövegekből kétirányú LSTM-mel
- Csapatnév: Bokor, Mészáros, Scheier
- Résztvevők: Mészáros Bálint, Scheier Balázs