A webaratásból származó szövegek automatikus feldolgozása

Simon Eszter

404 Not Found – Ki őrzi meg az internetet?

OSZK, 2023. 11. 29.

Az előadás vázlata

1. A webaratás eredménye
2. A nyelvfeldolgozásról általában
3. A webarchivált szövegek nyelvi elemzése
4. Az eredmény hasznosítása

A webaratás eredménye

A webaratás eredménye

3 eredményfájl:

- WARC: egy konténer, amiben benne van minden, ami le lett szedve
- WAT: metaadatfájl
- WET: tartalmazza a szöveget, de ebben még benne van minden, ami egy weboldalon van

további feladatok:

- boilerplate removal: nem kellő HTML-részek eltávolítása → pl. navigációs linkek, fejléc, lábléc

https://nlp.fi.muni.cz/projects/justext/

A nyelvfeldolgozásról általában

Természetesnyelv-feldolgozás és szinonimái

- természetesnyelv-feldolgozás (natural language processing, NLP)
- számítógépes nyelvészet (computational linguistics, CL)
- nyelvtechnológia (human language technology, HLT)
- mesterségesintelligencia-kutatás (Artificial Intelligence (AI) research

Alkalmazások

a nyelvtechnológiai fejlesztések tipikusan nagyobb alkalmazásokba beépítve jelennek meg

- helyesírás-ellenőrzés: böngészők, szerkesztők
- auto-complete
- természetes nyelvű keresés a böngészőben
- gépi fordítás: Google Translate, DeepL
- automatikus beszédgenerálás a GPS alkalmazásban
- diktálás írott szöveggé alakítása a mobilon
- személyi asszisztensek: Siri, Alexa, Cortana, Google Assistant
- chatbotok, ChatGPT
- szentimentelemzés (pozitív, negatív és semleges értékelések)
- Google Calendar bejegyzés e-mailekből

A nyelvtechnológia feladata

- bár a nyelvtechnológia folyamatosan fejlődik, még távolról sem tekinthető megoldottnak a nyelvfeldolgozás minden lépése ← az emberi nyelv komplexitása
- az egyik fő nehézség: az ember az értelmezés során számos nehezen formalizálható tényezőt is figyelembe vesz
  - a megnyilatkozás körülményei (hol, mikor, kikkel)
  - többletjelentés (ígéret, fenyegetés, irónia)
- a nyelvtechnológia feladata jelenleg: a szövegfolyamban detektálható releváns információ adott célnak megfelelő feldolgozása

A nyelvfeldolgozás lépései

1. a digitális adatfolyam automatikus feldolgozása
2. az eredeti anyagban expliciten nem szereplő információ megtalálása
3. az adatok strukturált formába szervezése
4. az eredményeknek a felhasználó számára optimális prezentálása

A szövegfeldolgozás részfeladatai

- mondatokra bontás
- szavakra bontás (tokenizálás)
- morfológiai elemzés
- morfológiai egyértelműsítés
- szintaktikai elemzés
- tulajdonnév-felismerés
- koreferenciafeloldás
- mondatok közötti összefüggések felismerése
- szemantikai relációk detektálása
- érzelmek detekciója

A webarchivált szövegek nyelvi elemzése

Mit csinál a nyelvi elemző?

bemenet:

a lecsupaszított folyó szöveg

lépések:

1. mondatra bontás
2. tokenizálás
3. morfológiai elemzés
4. morfológiai egyértelműsítés
5. (tulajdonnév-felismerés)
6. (szintaktikai elemzés)

Mondatra bontás és tokenizálás

mondatrabontás

- mondathatárok azonosítása
- minden mondat
- pontos problémák, egyéb nehézségek

tokenizálás

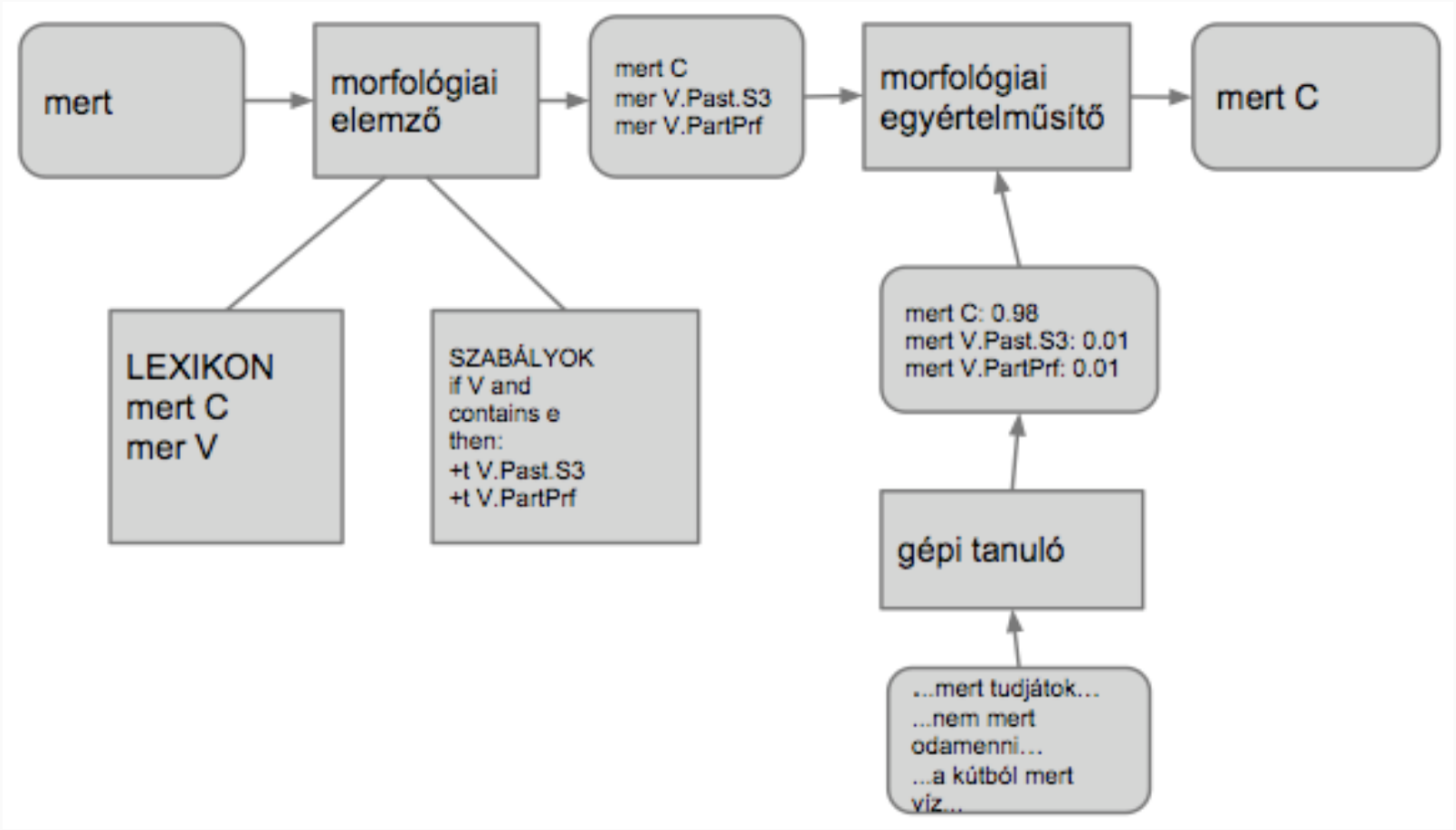
- szóalkotó karakterek és szónemalkotó karakterek
- számok, informatikai kifejezések, smiley-k...

Morfológiai elemzés

tokenszintű elemzés → nem lát se előre, se hátra → nincs kontextus → többértelműség

```
falucska fa[N]luc[N]ska[N]
falucska fa[N]lucsok[N]a[PxS3]
falucska falucsok[N]a[PxS3]
falucska falu[N]cska[_Dim=cskA]
falucska falucska[N]
```

Morfológiai egyértelműsítés



Tulajdonnév-felismerés

1. a nevek lokalizálása strukturálatlan szövegben
2. a megtalált elemek besorolása előre definiált névosztályokba

Szintaktikai elemzés

3 típusa:

- sekély szintaktikai elemzés (chunking): főnévi frázisok megtalása a mondatban
- összetevős elemzés: azt tárja fel, hogy a szavak egymással kombinálódva milyen kifejezéseket alkotnak, illetve hogyan állnak össze egy mondattá
- függőségi elemzés: a mondatok szerkezeti egységei közötti függőségi viszonyokat (pl. alany, tárgy, jelző) tárja fel

https://e-magyar.hu/hu/parser
https://huggingface.co/spaces/huspacy/demo

További lépések

- A nyelvileg elemzett szövegen statisztikát számol.
- Kiszűri
  - a nem kellő szófajú szavakat,
  - a punktuációkat,
  - a space jellegűeket,
  - a csak számokat tartalmazó tokeneket,
  - az egy kisbetűből álló tokeneket,
  - az üres sorokat és
  - a stopszólistában szereplő szavakat.
- A morfológiai kódokat lecsupaszítja puszta szófajkódokra.
- Sorba rendez, és gyakoriságot számol.
- A maradékot gyakorisági alapon sorba rendezi.
- Levágja a 2000 leggyakoribbat.
- A bemeneti fájl nevéből kiparszolja a dátumot, és egy külön oszlopba beleilleszti a kimenő fájlba.

Az eredmény hasznosítása

Mit nyerünk a szövegfeldolgozással?

- lesz egy mondatokra és szavakra bontott dokumentumunk:
  - szó- és mondatstatisztikák → szerzőazonosítás, stilometria
- minden szónak tudni fogjuk a tövét:
  - tőalapú keresés vs. szóalakalapú keresés → okosabb keresés, több találat
  - tőalapú statisztikák → szókincs
- minden szónak tudni fogjuk a szófaját:
  - szófajalapú statisztikák
- ismerni fogjuk a szövegben szereplő neveket:
  - megismerjük a szereplőket, helyszíneket stb.
  - térképre rakhatjuk őket
  - lehorgonyozhatjuk őket különféle adatbázisokhoz, névterekhez
- kibányászhatjuk a szövegbeli információkat, eseményeket
  - ki, hol, mikor, mit csinált?

Magyar hírportálok orosz–ukrán háborús tartalmai

- 2022. február 21-étől
- a hazai és a határon túli online sajtóban megjelenő ukrajnai hírek mentése → eseményalapú webaratás
- az ukrán háborúval foglalkozó online magyar hírek szóhasználatának változását feltáró kutatás, illetve annak vizualizációja: *https://dhupla.hu/page/kreativ/ukrajna-hirek-szokeszlet*

Tárgyszavazás

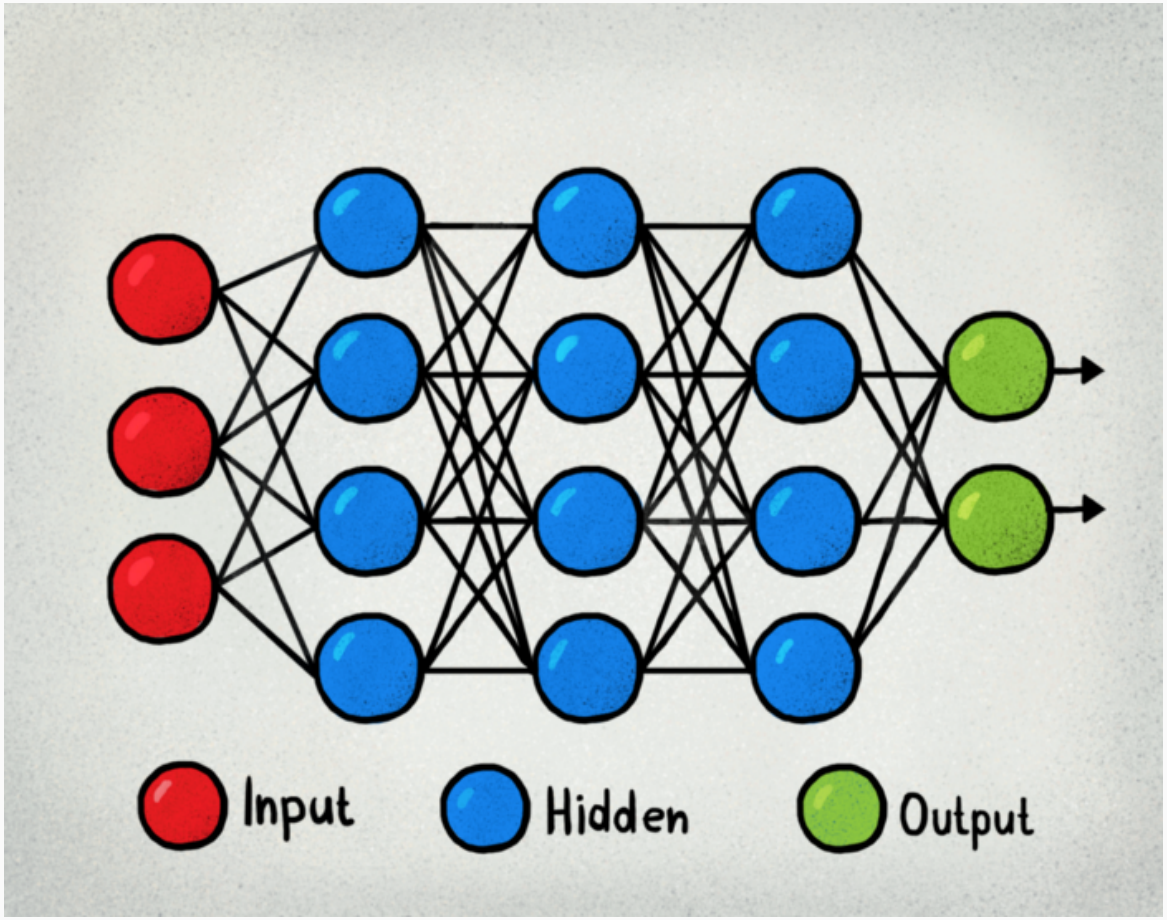
- tárgyszavak listájából választunk, azokkal címkézzük fel a dokumentumokat
- a webarchivált fájlok egy része rendelkezik kézzel hozzáadott tárgyszavakkal és témakörökkel
- a felcímkézett dokumentumok tanító adathalmazként szolgálhatnak egy gépi tanuló algoritmus számára → **felügyelt gépi tanulás (supervised learning)**

Álhírek automatikus azonosítása

- együttműködés a Szegedi Tudományegyetemmel
- „Álhírek, áltudományos nézetek nyelvészeti azonosítása” című projekt → az MTA által támogatott *Tudomány a Magyar Nyelvért Nemzeti Program* 4. alprogramja
- a webarchívum egyészségügyi tematikus részgyűjteménye
- automatikus álhírfelismerő fejlesztése
- nyelvészeti módszerekkel feltárt konteó: azon elemeknek a kimutatása, amelyek az álhíreket jellemzik

Nagy nyelvi modellek (Large Language Models (LLMs))

mesterséges neurális hálók, mélytanulás



GPT: Generative Pre-trained Transformer

- előtanított nyers nyelvmodell
- szöveggenerálásra képezték
- a következő szó kitalálása → a tanítókorpuszban látott adatok alapján a bal oldali előzményből kell megjósolnia a legvalószínűbb folytatást
- nagyon nagy mennyiségű tanítóadatot igényel
- GPT-3 (2020): több mint 175 milliárd paraméter

együttműködés a Nyelvtudományi Kutatóközponttal:

a webarchivált szövegek tanítóanyagként tudnak szolgálni a magyar GPT, a PuliGPT számára:
https://juniper.nytud.hu/demo/puli

Köszönöm a figyelmet!

simon.eszterke@gmail.com