Bevezetés

Nyelvtechnológia olvasószeminárium – 2019/20 tavasz

Simon Eszter

MTA Nyelvtudományi Intézet

Tartalom

Bemutatkozás

Bemutatkozás

- mi
- ti

A félév bemutatása

Tanrend

Szorgalmi időszak

Első nap: 2020. február 10. (hétfő)
Tavaszi szünet: 2020. április 6. – április 17. (hétfő–péntek)
Ünnepnap: 2020. május 1. (péntek)
Utolsó tanítási nap: 2020. május 15. (péntek)

Vizsgaidőszak

Első nap: 2020. május 18. (hétfő)
Utolsó nap: 2020. június 27. (szombat)

Az órák

- összesen 14 hét
- ebből 3 elmarad
- összesen 11 óra
- ha kell, csinálhatunk pótlást
- olvasószeminárium

Nyelvtechnológia tételsor

1. Az n-gramok mibenléte. n-gram-alapú statisztikai megfigyelések. Az n-gramok használata a nyelvtechnológia különböző területein.
2. Véges automaták, véges fordítók és kiterjesztéseik. Reguláris kifejezések. Műveletek automatákkal. Az elemzés és a generálás hasonlósága és különbsége (a nyelvi jelenségek és a lexikon szerepe). A kétszintes leírás alapelvei.
3. Természetes nyelvek mondatszerkezetének ábrázolása formális nyelvtanokkal. Függőség és összetevős szerkezet: hasonlóságok és különbségek. Szintaktikai elemzési algoritmusok. A gépi és az emberi mondatelemzés összevetése.

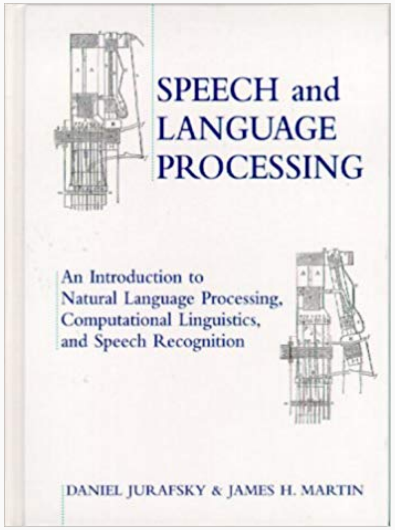
Nyelvtechnológia tételsor – folyt.

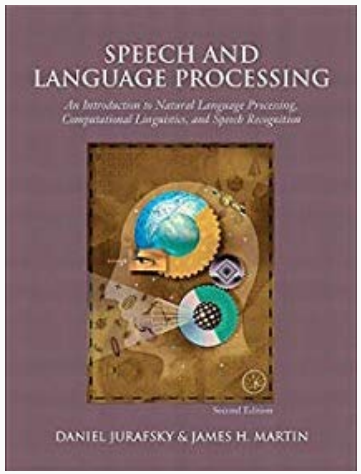
4. Az unifikáció és az unifikálhatóság fogalma. Jegyszerkezetek és kapcsolatuk az irányított körmentes gráfokkal. Példák az unifikáció alkalmazására a morfológiában és a szintaxisban.
5. Lexikai és mondatjelentés-reprezentációk. Fogalmi hálók, ontológiák. A WordNet és továbbfejlesztései. Szóbeágyazási modellek.
6. A gépi fordítás alapvető módszerei. Szabály-alapú közelítések. Párhuzamos korpuszok és felhasználásuk. A statisztikai gépi fordítás alapjai. A neurális fordítás alapgondolata.

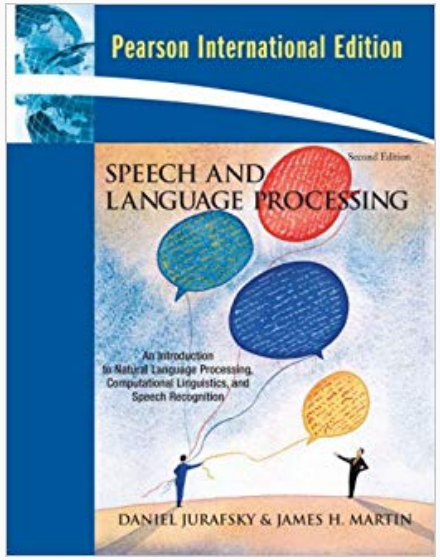
Olvasnivaló

Dan Jurafsky – James H. Martin: Speech and Language Processing

3rd edition draft: https://web.stanford.edu/~jurafsky/slp3/
2nd edition: https://readyforai.com/download/speech-and-language-processing-2nd-edition-pdf/
1st edition







Tematika

Google doksi: https://docs.google.com/document/d/1CUTOuH1OhK4XKaO65VbK8kao5vilR1RKzgNHvEvXsNc/edit?usp=sharing

GitHub repo: https://github.com/ferenczizsani/nyelvtech

Fejezetek

1. Language Modeling with N-Grams (3rd edition Chapter 3)
2. Regular Expressions and Automata (2nd edition Chapter 2 = 1st edition Chapter 2)
3. Words and Transducers (2nd edition Chapter 3)
4. Constituency Grammars (3rd edition Chapter 12)
5. Constituency Parsing (3rd edition Chapter 13)
6. Statistical Constituency Parsing (3rd edition Chapter 14)
7. Dependency Parsing (3rd edition Chapter 15)
8. Features and Unification (2nd edition Chapter 15 = 1st edition Chapter 11)
9. Logical Representations of Sentence Meaning/Representing Meaning (3rd edition Chapter 16 = 2nd edition Chapter 17)
10. Word Senses and WordNet (3rd edition Chapter 19)
11. Vector Semantics and Embeddings (3rd edition Chapter 6)
12. Machine Translation (2nd edition Chapter 25)

Teljesítés

Jelenléti követelmények: Legfeljebb 3 óráról lehet hiányozni – ez és az órai aktivitás az aláírás feltételei.

Félévközi teljesítés: Fejezeteket kell elolvasni, és azokat a többieknek prezentálni (handout, prezentáció stb.).

A félév végi osztályzat: A félévközi teljesítés alapján.

Konzultáció: E-mailben egyeztetett időben (simon.eszter@nytud.mta.hu).