Word Senses and WordNet

Ferenczi Zsanett

2020. május 8.

Vázlat

1. Bevezetés
2. Jelentés, szemantikai relációk
3. WordNet
4. Jelentés-egyértelműsítés
5. Szóbeágyazások fejlesztése
6. Word Sense Induction

Bevezetés

Bevezetés

- a szavak sokszor bírnak több jelentéssel → homonim szavak
  (pl. angol **bank**)
  (ha van kapcsolat a jelentések között: poliszémia, pl. **egér**)
- kérdésmegválaszolásnál, gépi fordításnál fontos a jelentések elkülönítése (pl. **bat** → **denevér / baseball ütő**) → word sense disamiguation
- probléma: szóbeágyazásokkal az antonimák közel kerülnek egymáshoz: tezauruszok segíthetnek (pl. **up, down**)
- WordNet: online tezaurusz, szemantikai relációkat tartalmazó adatbázis

Jelentés, szemantikai relációk

Szójelentés

- `szemantikai relációk`: pl. szinonímia, antonímia, `is-a` reláció, WordNeten ezek fel vannak sorolva
- mikor beszélhetünk különálló jelentésekről?
  - ha egymástól független igazságfeltételük van
  - ha eltérő szintaktikai viselkedést mutatnak
  - ha egymástól független szemantikai relációik van
  - ha ellentétes értelműek
- `zeugma`: egy szó két használatát összekapcsoljuk egy mondaton belül, ellentmondó, furcsa olvasat

  `?Does Lufthansa serve breakfast and San Jose?`

Szemantikai relációk

- `szinonímia`: két lemma jelentése (közel) azonos (pl. **`couch`** és **`sofa`**)
- `antonímia`: ellentétes jelentések (pl. **`long`** és **`short`**)
- `hiponímia`: X hiponimája Y-nak (= Y hipernimája X-nek), ha X specifikusabb, Y alosztálya, ha X implikálja Y-t
  (pl. **`kutya`** $\rightarrow$ **`állat`**)
- `meronímia`: rész-egész viszony (pl. **`kerék`** meronimája a **`kocsinak`**, **`kocsi`** holonimája a **`keréknek`**)
- `metonímia`: strukturált poliszémia, egy szó más jelentést vesz fel valamilyen ok-okozati, vagy térbeli, időbeli érintkezésen keresztül (pl. **`bank`**: épület és pénzügyi intézet)
  `épület <=> intézet`

WordNet

WordNet

- lexikális szemantikai hálózat
- angol WordNet három adatbázisból áll: egy a főneveknek, egy az igéknek, és egy a mellékneveknek és határozószóknak
- szavak gyűjteménye, mindegyik annotálva jelentések halmazával
- `WordNet 3.0`: 117 798 főnév, 11 529 ige, 22 479 melléknév, és 4 481 határozószó
- átlagosan egy főnévnek 1,23 jelentése van, egy igének 2,16
- nincs jelölve, ha más kiejtés társul más jelentésekhez
- `synset`: közeli szinonimák halmaza, ezek határoznak meg egy-egy fogalmat, és ezek állnak kapcsolatban más synsetekkel

WordNet példa

Adjective

- S: (adj) **cold** (having a low or inadequate temperature or feeling a sensation of coldness or having been made cold by e.g. ice or refrigeration) *"a cold climate"; "a cold room"; "dinner has gotten cold"; "cold fingers"; "if you are cold, turn up the heat"; "a cold beer"*
  - *see also*
  - *similar to*
  - *attribute*
  - *antonym*
    - W: (adj) hot [Opposed to: cold] (used of physical heat; having a high or higher than desirable temperature or giving off heat or feeling or causing a sensation of heat or burning) *"hot stove"; "hot water"; "a hot August day"; "a hot stuffy room"; "she's hot and tired"; "a hot forehead"*
  - *derivationally related form*

- S: (n) **bat**, chiropteran (nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *member holonym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
      - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
        - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
          - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
            - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
              - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)

WordNet

The noun "bass" has 8 senses in WordNet.

1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

- logikai term-ek helyett szavak (jelentések) listájával határoz meg egy-egy fogalmat
- `{chump[1], fool[2], gull[1], mark[9], patsy[1], fall guy[1], sucker[1], soft touch[1], mug[2]}`
- **definíció:** `a person who is gullible and easy to take advantage of`

Szemantikai kategóriák

- `supersense`: szemantikai kategóriák, nagyobb osztályokba sorolja a synseteket
- főnevek: 26 kategória
- igék: 15 kategória

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

Szemantikai relációk a WordNetben

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} $\rightarrow$ *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} $\rightarrow$ *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} $\rightarrow$ *author*^{1}
Instance Hyponym	Has-Instance	From concepts to their instances	*composer*^{1} $\rightarrow$ *Bach*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} $\rightarrow$ *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} $\rightarrow$ *meal*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} $\Longleftrightarrow$ *follower*^{1}
Derivation		Lemmas w/same morphological root	*destruction*^{1} $\Longleftrightarrow$ *destroy*^{1}

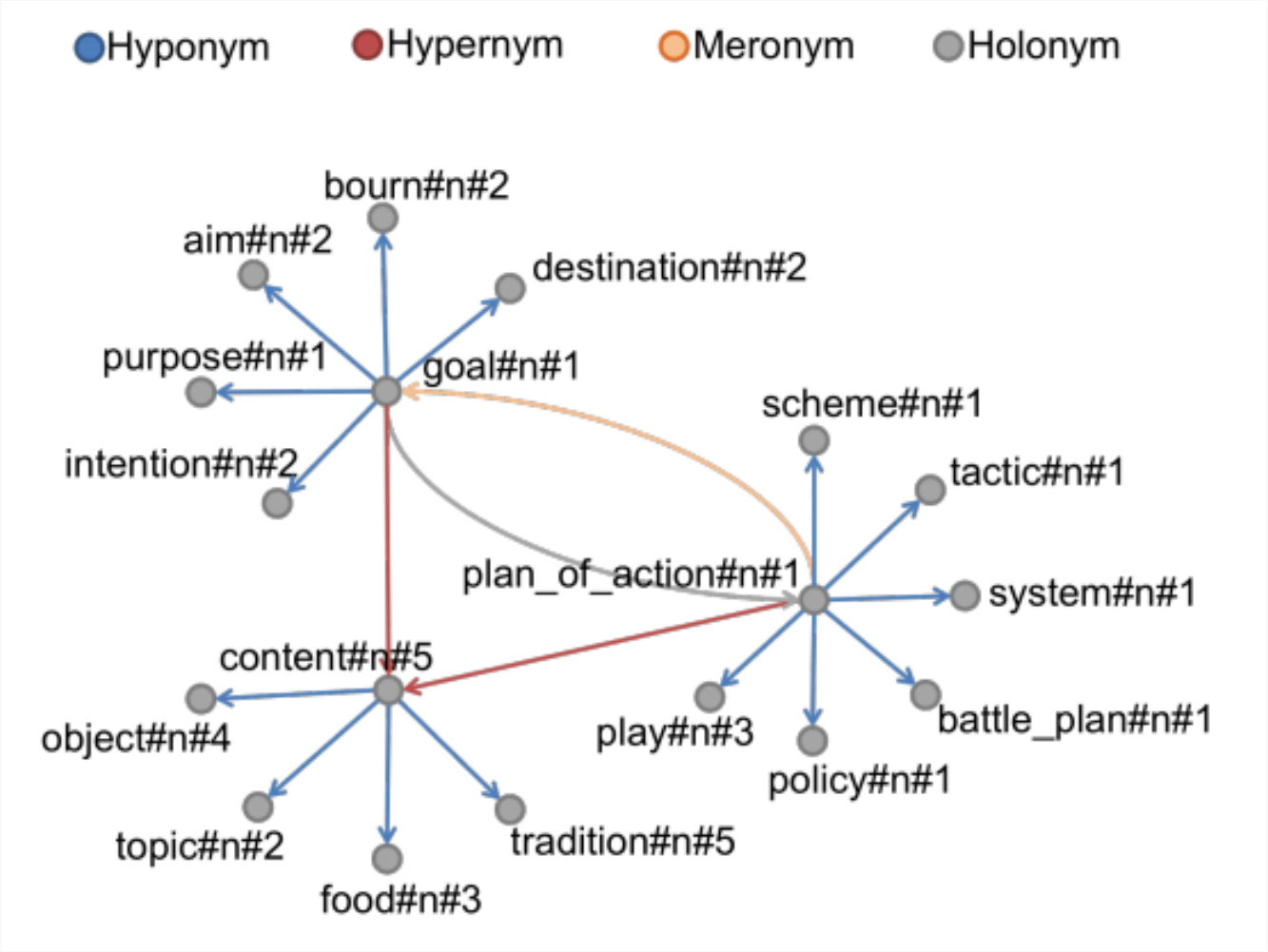
Figure 19.3 Some of the noun relations in WordNet.

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly*^{9} $\rightarrow$ *travel*^{5}
Troponym	From events to subordinate event	*walk*^{1} $\rightarrow$ *stroll*^{1}
Entails	From verbs (events) to the verbs (events) they entail	*snore*^{1} $\rightarrow$ *sleep*^{1}
Antonym	Semantic opposition between lemmas	*increase*^{1} $\Longleftrightarrow$ *decrease*^{1}

Figure 19.4 Some verb relations in WordNet.

- `is-a` reláció (hypernym)
- `has-a` reláció (hyponym)

Szemantikai relációk a WordNetben



Chaplot és Salakhutdinov (2018)

Taxonómia

```
bass[7] (member with the lowest range of a family of instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

- kétféle taxonómiai entitás létezik: `osztály` és `példány`
- `példányok`: tulajdonnevek, egyedi entitások, nincs hiponimájuk
- pl. `Beethoven`: német zeneszerző → `példány` vs. Beethoven zenéje → `osztály`
- Magyar WordNet (HuWN):
  - 50 000 szó, 60 000 jelentés, 42 000 synset
  - HuWN

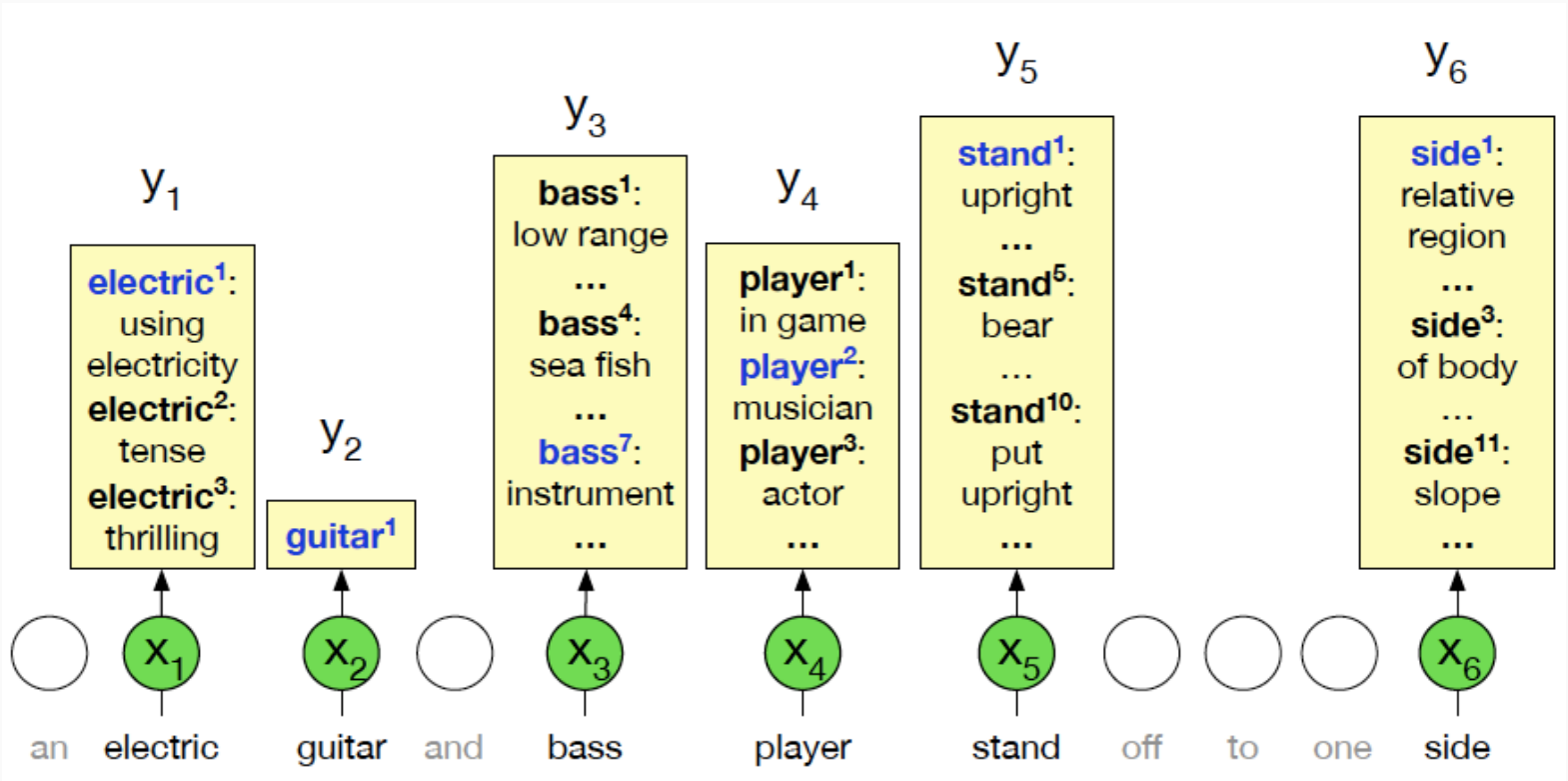
Jelentés-egyértelműsítés

Jelentés-egyértelműsítés

- `word sense disambiguation`
- egy jelentés kiválasztása egy adott szóalakhoz egy adott jelentéshalmazból (`inventory of sense tags`)
- pl. WordNetet használva, vagy annak nagyobb kategóriáit
- `WSD` két iránya:
  - csak egyes szavak (`lexical sample`): szavak és jelentések halmaza kicsi, felügyelt osztályozási modellek
  - minden szövegbeli szó (`all-word`): betanítás semantic concordance-ból: olyan korpusz, melyben minden (nyílt szóosztálybeli) szó fel van címkézve
    pl. **`SemCor`** korpusz

SemCor és all-word

You will find$^{9}_{v}$ that avocado$^{1}_{n}$ is$^{1}_{v}$ unlike$^{1}_{j}$ other$^{1}_{j}$ fruit$^{1}_{n}$ you have ever$^{1}_{r}$ tasted$^{2}_{v}$



Kontextuális vektor

- egy szót saját kontextusában vizsgálunk
- adott jelentéshez rendel vektort (egy szóalakhoz több különböző vektor is tartozhat, kontextustól függően)
- fontos a szavak sorrendje
- szükség van a mélytanulásra, több rejtett rétegre
- modellek: `ELMo` és `BERT`

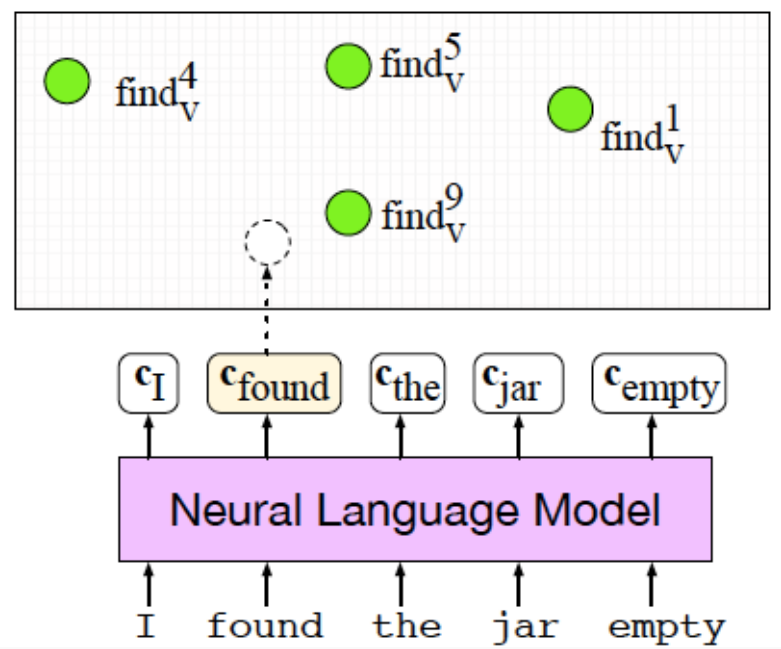
- `He went to the prison cell with his cell phone to extract blood cell samples from inmates.`

Jelentés-egyértelműsítés

- `leggyakoribb jelentés`
  - a WordNetben gyakoriság szerint vannak rendezve az egyes jelentések
  - az 1. jelentés a leggyakoribb
- `azonos jelentés`
  - egy tokenhez egy diskurzuson belül ugyanazt a jelentést rendeljük
  - homonimák esetén hasznosabb
  - poliszém szavaknál kevésbé működik

Algoritmusok

- legközelebbi szomszéd algoritmus (`1-nearest-neighbor`)
- ha nem találjuk a tanítóanyagban a szót:
  - vehetjük a leggyakoribb jelentést
  - a WordNetben alulról felfelé haladva visszalépegetünk



További lehetőségek

- `feature-based WSD` (jegyalapú):
  - a célszó körüli szavak szófaji címkéi (pl. 2-2 szóhoz mindkét oldalon)
  - kollokációs jegyek (szomszédos szó n-grammok)
  - súlyozott szóbeágyazási átlag (pl. 2-2 szomszédos szó beágyazásának átlaga)
- tudásalapú algoritmusok: pl. Lesk algoritmus:
  - csak WordNetet (és hasonló erőforrásokat) használ
  - `Simplified Lesk`
  - azt a jelentést keresi, amelyben a legtöbb azon szavak száma, amely a kontextusban is szerepel
  - `idf`-fel ki lehet egészíteni
  - szóbeágyazások koszinusza a közös szavak száma helyett jobb eredményt ad

Word-in-Context kiértékelés

- `Word-in-Context` kiértékelés
  - Word-in-Context dataset: egy mondatpár, bennük ugyanaz a célszó
  - el kell dönteni, hogy azonos vagy különböző jelentésben szerepelnek-e
  - főleg WordNetből vett példamondatok
  - ennek egy megoldása: mindkét célszóhoz kiszámoljuk a kontextuális beágyazást, kiszámoljuk a koszinuszt, és ha egy adott küszöbérték feletti, akkor azonos jelentésűek

F	There's a lot of trash on the **bed** of the river —
	I keep a glass of water next to my **bed** when I sleep
F	**Justify** the margins — The end **justifies** the means
T	**Air** pollution — Open a window and let in some **air**
T	The expanded **window** will give us time to catch the thieves —
	You have a two-hour **window** of clear weather to finish working on the lawn

Wikipedia

- WSD-ben Wikipedia is használható olyan adathalmazként, amelyben a jelentések fel vannak címkézve
- a Wikipedia szócikkeknek van egyedi azonosítója
- pl. `[[bar (law)|bar]]`, `[[bar (music)|bar]]`
- le kell képezni ezeket a fogalmakat arra, amit használni akarunk (pl. WordNet) (`mappelés`)
- `BabelNet` (jelentés-annotált erőforrás):
  Wikipediából készült, WordNet synsetekre van leképezve többnyelvű erőforrás

Szóbeágyazások fejlesztése

Szóbeágyazások fejlesztése

- antonimák szóbeágyazása sokszor nagyon hasonló
- a relációkat a tezauruszból beleépíthetjük a szóbeágyazások tanulásába
- szóbeágyazások tanulása után a szinonimákat közelebb, antonimákat egymástól távolabbra kényszerítjük $\rightarrow$ `retrofitting, counterfitting` metódus

	Before counterfitting			After counterfitting		
east	west	north	south	eastward	eastern	easterly
expensive	pricey	cheaper	costly	costly	pricy	overpriced
British	American	Australian	Britain	Brits	London	BBC

Word Sense Induction

Word Sense Induction

- `Word Sense Induction / Discrimination`
- nehéz nagy, felcímkézett korpuszt találni/csinálni
- egy felügyelet nélküli megoldás: `WSI`
- nincsenek emberek által meghatározott szójelentések
- tanítás:
  - kontextus vektor kiszámolása minden tokenre
  - klaszterező algoritmus egy meghatározott számú klaszterbe rendezi ezeket
  - minden klaszterre számoljunk átlagot (`centroid`)
  - ezek az átlagolt vektorok lesznek az adott jelentést reprezentáló vektorok
- ahhoz, hogy megállapítsuk `w` szó egy `t` tokenjének jelentését:
  - számoljuk ki `t` kontextus vektorát
  - minden jelentésvektort nyerjünk ki a `w` szóhoz
  - `t` jelentése: ezek közül a hozzá legközelebbi

Felügyelet nélküli módszerek kiértékelése

- `extrinsic`
  - beépítve egy nagyobb alkalmazásba
- `intrinsic`
  - automatikusan kinyert jelentésosztályok leképezése egy kézzel készült, gold standard halmazra → össze tudjuk hasonlítani őket
  - több megoldás született már (pl. klaszterek közötti egyezés)
  - nincs még standard kiértékelési metrika

Bibliográfia

Devendra Singh Chaplot and Ruslan Salakhutdinov. Knowledge-based word sense disambiguation using topic models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 3.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. Methods and Results of the Hungarian WordNet Project. In *Proceedings of The Fourth Global WordNet Conference*, pp. 311–321, 2008.