

# WORD SENSES AND WORDNET

---

Ferenczi Zsanett

2020. május 4.

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

# Bevezetés

---

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

## Jelentés, szemantikai relációk

---

- **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?

- **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** → **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

---



- 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

## 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)

The noun “bass” has 8 senses in WordNet.

1. bass<sup>1</sup> - (the lowest part of the musical range)
2. bass<sup>2</sup>, bass part<sup>1</sup> - (the lowest part in polyphonic music)
3. bass<sup>3</sup>, basso<sup>1</sup> - (an adult male singer with the lowest voice)
4. sea bass<sup>1</sup>, bass<sup>4</sup> - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass<sup>1</sup>, bass<sup>5</sup> - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass<sup>6</sup>, bass voice<sup>1</sup>, basso<sup>2</sup> - (the lowest adult male singing voice)
7. bass<sup>7</sup> - (the member with the lowest range of a family of musical instruments)
8. bass<sup>8</sup> - (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<sup>1</sup>, fool<sup>2</sup>, gull<sup>1</sup>, mark<sup>9</sup>, patsy<sup>1</sup>, fall guy<sup>1</sup>, sucker<sup>1</sup>, soft touch<sup>1</sup>, mug<sup>2</sup>}
- 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	<i>service</i>	GROUP	<i>place</i>	PLANT	<i>tree</i>
ANIMAL	<i>dog</i>	LOCATION	<i>area</i>	POSSESSION	<i>price</i>
ARTIFACT	<i>car</i>	MOTIVE	<i>reason</i>	PROCESS	<i>process</i>
ATTRIBUTE	<i>quality</i>	NATURAL EVENT	<i>experience</i>	QUANTITY	<i>amount</i>
BODY	<i>hair</i>	NATURAL OBJECT	<i>flower</i>	RELATION	<i>portion</i>
COGNITION	<i>way</i>	OTHER	<i>stuff</i>	SHAPE	<i>square</i>
COMMUNICATION	<i>review</i>	PERSON	<i>people</i>	STATE	<i>pain</i>
FEELING	<i>discomfort</i>	PHENOMENON	<i>result</i>	SUBSTANCE	<i>oil</i>
FOOD	<i>food</i>			TIME	<i>day</i>

**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	<i>breakfast</i> <sup>1</sup> → <i>meal</i> <sup>1</sup>
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> <sup>1</sup> → <i>lunch</i> <sup>1</sup>
Instance Hypernym	Instance	From instances to their concepts	<i>Austen</i> <sup>1</sup> → <i>author</i> <sup>1</sup>
Instance Hyponym	Has-Instance	From concepts to their instances	<i>composer</i> <sup>1</sup> → <i>Bach</i> <sup>1</sup>
Part Meronym	Has-Part	From wholes to parts	<i>table</i> <sup>2</sup> → <i>leg</i> <sup>3</sup>
Part Holonym	Part-Of	From parts to wholes	<i>course</i> <sup>7</sup> → <i>meal</i> <sup>1</sup>
Antonym		Semantic opposition between lemmas	<i>leader</i> <sup>1</sup> ⇔ <i>follower</i> <sup>1</sup>
Derivation		Lemmas w/same morphological root	<i>destruction</i> <sup>1</sup> ⇔ <i>destroy</i> <sup>1</sup>

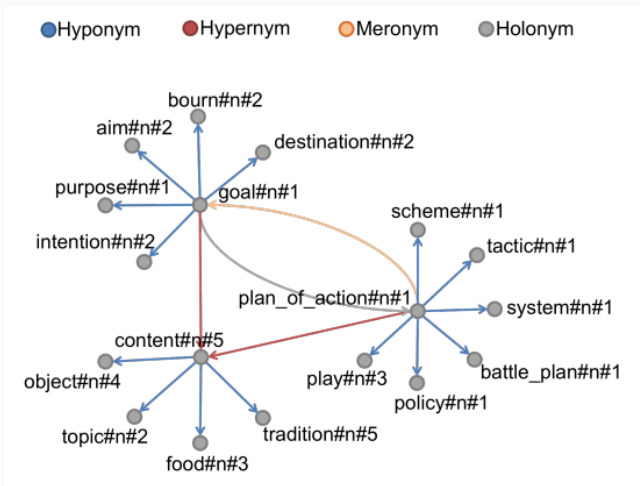
**Figure 19.3** Some of the noun relations in WordNet.

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> <sup>9</sup> → <i>travel</i> <sup>5</sup>
Troponym	From events to subordinate event	<i>walk</i> <sup>1</sup> → <i>stroll</i> <sup>1</sup>
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> <sup>1</sup> → <i>sleep</i> <sup>1</sup>
Antonym	Semantic opposition between lemmas	<i>increase</i> <sup>1</sup> ⇔ <i>decrease</i> <sup>1</sup>

**Figure 19.4** Some verb relations in WordNet.

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

# Szemantikai relációk a WordNetben



Chaplot és Salakhutdinov (2018)

```
bass7 (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

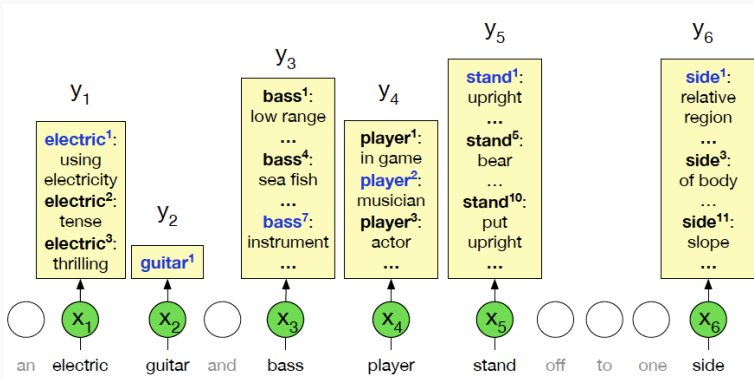
---



- **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<sub>v</sub><sup>9</sup> that avocado<sub>n</sub><sup>1</sup> is<sub>v</sub><sup>1</sup> unlike<sub>j</sub><sup>1</sup> other<sub>j</sub><sup>1</sup> fruit<sub>n</sub><sup>1</sup> you have ever<sub>r</sub><sup>1</sup> tasted<sub>v</sub><sup>2</sup>

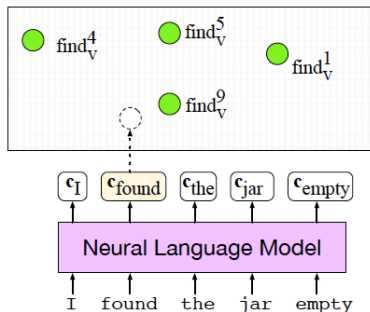


- egy szót saját kontextusában vizsgálunk
  - adott jelentéshez rendel vektort (egy szóalakhoz több különböző vektor is tartozhat, kontextustól függően)
  - fontos a szavak sorrendje
  - szükség van a mélytanulásra, több rejtett rétegre
  - modellek: ELMo és BERT
- 
- He went to the prison cell with his cell phone to extract blood cell samples from inmates.

- **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



- **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 dataset: egy mondatpár, bennük ugyanaz a célszó
  - el kell dönteneti, 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

- 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 tezausból beleépíthetjük a szóbeágyazások tanulásába
- szóbeágyazások tanulása után a szinonimákat közelebb, antonimákat egymástól távolabbra kényszerítjük → **retrofitting, counterfitting** metódus

Before counterfitting				After counterfitting		
<b>east</b>	west	north	south	eastward	eastern	easterly
<b>expensive</b>	pricey	cheaper	costly	costly	pricy	overpriced
<b>British</b>	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

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

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