A nyelvtechnológia területei

korpuszépítés, tulajdonnevek, koreferencia

Vadász Noémi

Számítógépes nyelvészet
Eötvös Loránd Tudományegyetem
2021/2022 tavasz

Egy hosszú bevezető: mire is kell a korpusz?

Módszerek a számítógépes nyelvészetben

- szabályalapú
- gépi tanulás
  - felügyelt
  - félig felügyelt
  - felügyeletlen
- neurális gépi tanulás

Erre kellenek a korpuszok:

- szabályalapú megoldások kiértékelésére
- gépi tanulási és neurális gépi tanulási megoldások tanítására és kiértékelésére

Szabályalapú

☺ a fejlesztőnek nagy kontrollja van a rendszer fölött
☺ könnyen értelmezhető visszacsatolás
☺ magas pontosság
☺ nyelvi adatok, amik könnyen megragadhatók szabályokkal (reguláris kifejezésekkel), pl. dátumok szerkezete

☹ sok kézimunka, nagy szakértelem kell hozzá
☹ nem hibatűrő
☹ bonyolult a fejlesztése, törékeny
☹ nehezen átvihető más doménre, nyelvre
☹ lehetetlen olyan szabályrendszert írni, ami mindent lefed, amit kell, de semmit, amit nem
☹ a fedés a listák és a szabályok számának növelésével javítható, de a szabályok száma, a lexikon mérete korlátozott

- pl. morfológiai elemzés, tokenizálás

Statisztikai (sztochasztikus), klasszikus gépi tanulás

- adatorientált, gyakorisági adatokból indul ki
- a nyelv általánosabb megértése, modellálása
- kézzel kinyert jegyekre (feature) támaszkodik (pl. mondathossz, POS-tagek, spec. szavak előfordulása)
- gépi tanuló algoritmusok (pl. Naive Bayes, SWM, döntési fa stb.)
- nehézség: az adatok ritkasága („rare words are very common”)
- pl. szekvenciális címkézési feladatok, szintaktikai elemzés

Szekvenciális címkézés: NER

[...] közölte Wolf László, az OTP Bank vezérigazgató-helyettese az MTI érdeklődésére.

közölte	0
Wolf	B-PER
László	E-PER
,	0
az	0
OTP	B-ORG
Bank	E-ORG
vezérigazgató-helyettese	0
az	0
MTI	1-ORG
érdeklődésére	0
.	0

Felügyelt gépi tanulás

- feltételezés 1.: az adatpontok egymástól független elemek, amelyeknek egyenletes az eloszlásuk
- feltételezés 2.: az eddig nem látott adatpontokra is igaz a fenti állítás
- $\rightarrow$ a már látott nyelvi elemekből tudunk következtetni a még nem látottakra
- a nyelvtechnológiában: az annotált korpuszból tanulja ki a számítógép az adatpontokra jellemző jegyeket (majd annotált korpuszon is értékelünk ki)

A felügyelt gépi tanulás forgatókönyve (címkézési feladat)

1. gold standard korpusz
2. train-devel-test halmaz
3. jegykinyerés
4. modellépítés
5. predikálás (taggelés)
6. kiértékelés

train-devel-test

- a gold standard korpuszt felosztjuk halmazokra
  1. train: ezen tanítunk
  2. development: ezen fejlesztünk
  3. test: ezen értékelünk ki
- a teszthalmaz elemei nem szerepelhetnek a tanítóhalmazban!
- ez egyrészt csalás, másrészt a rendszer túlzottan ‘rátanul’ a szövegre, nem lesz képes az általánosításra

Jegykinyerés (feature extraction)

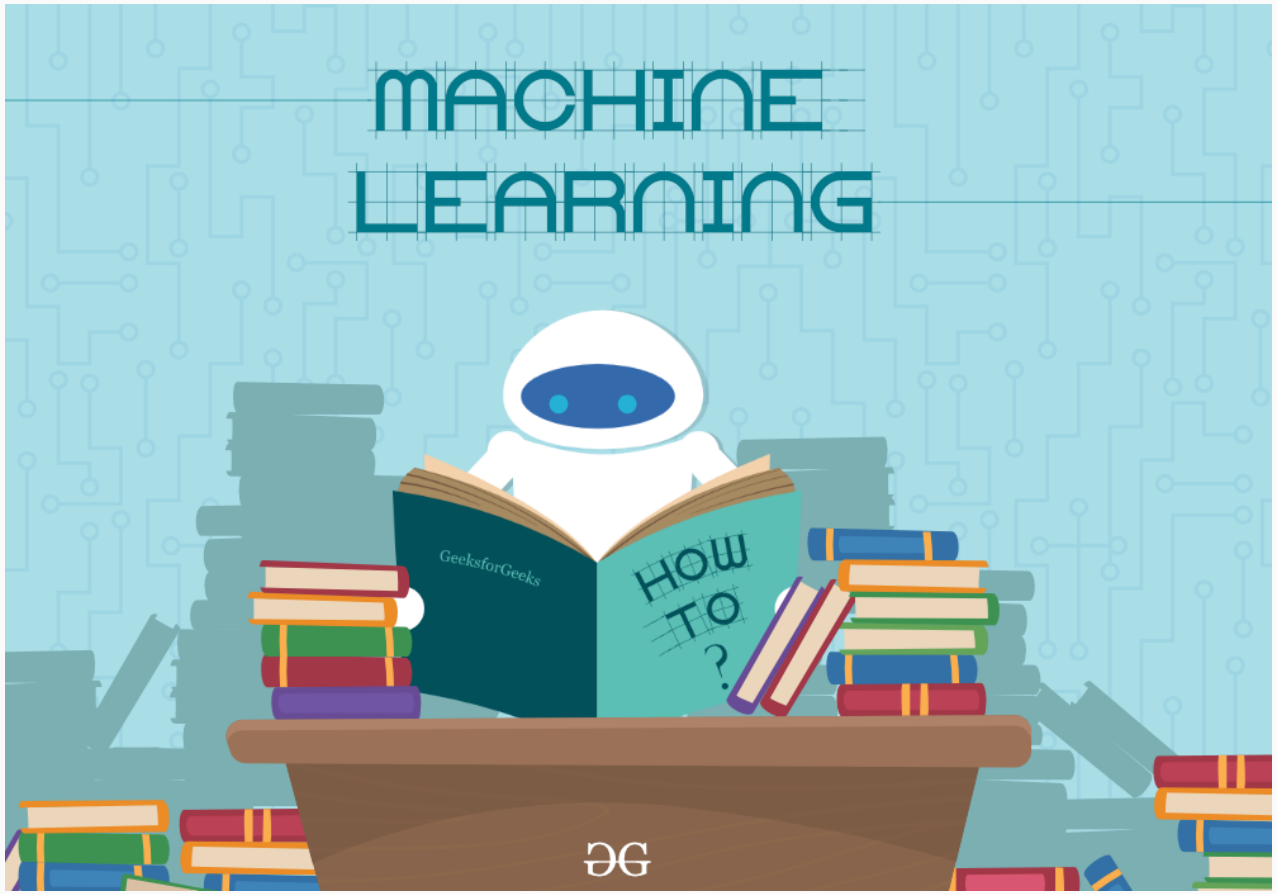
- a jegyek az adatpontok különféle tulajdonságait írják le
- a jegyeket a számítógépes nyelvész találja ki, definiálja és kódolja
- a jegy hasznosságát az adat határozza meg: a jegy megkülönböztető erejét ki kell mérni, utána eldönteni, hogy alkalmazzuk-e
- a jegyek hozzáadása vagy a paraméterek állítása egyesével, majd mérés, ha nem ront, akkor meghagyjuk
- a jegyek vektorokra képeződnek le

Jegyek a névelemfelismerésben

- Ortográfiai jellemzők
  kezdőbetű típusa, szóhossz, tartalmaz számot/írásjelet, arab/római szám
- Gyakorisági adatok
  kis/nagybetűs-, mondatközi nagybetűs/nagybetűs arányok, gyakoriság
- Szövegkörnyezet info trigger uni-/bi-/trigramok, mondatpozíció, dokumentumon belüli pozíció
- Kifejezésszintű információ
  megelőző tokenek címkéi, zárójelben/idézőjelben van, reguláris kifejezések
- Egyértelmű szavak szótára
  tanuló adatbázisból összegyűjtve, pl. betegségek nevei
- Trigger szótárak
  keresztnevek, országok, városok...

Modellépítés

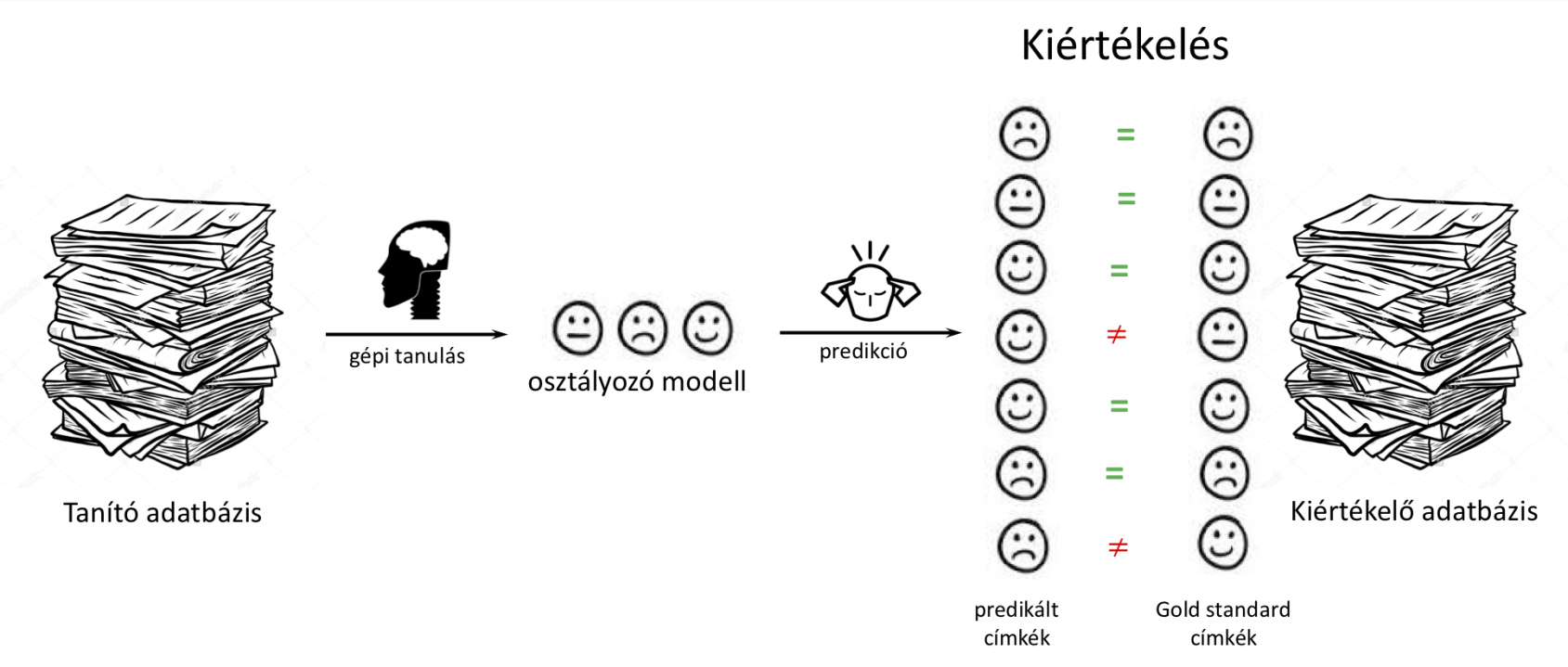
- a jegy-címke párokhoz súly van hozzárendelve, ami azt mutatja meg, hogy az adott jegy mennyire van hatással arra, hogy az adott jeggyel rendelkező token az adott címkét kapja



Predikálás és kiértékelés

- a teszthalmazon!
- a teszthalmaz feature-izálása, majd a feature-vektorok alapján a címkék kibocsátása
- az egyes tokenekhez azok a címkék kerülnek kiosztásra, amik a jegyvektorok alapján a legnagyobb valószínűséget kapták
- az eredményt összevetjük a teszthalmaz gold-standard címkéivel
- kiértékelés: pontosságot, fedést, F-mértéket számolunk
- a rendszerünk készen áll arra, hogy további szövegeket címkézzünk vele (várható pontossággal, fedéssel) :)

Honnan tudhatjuk, hogy egy eszköz jó?



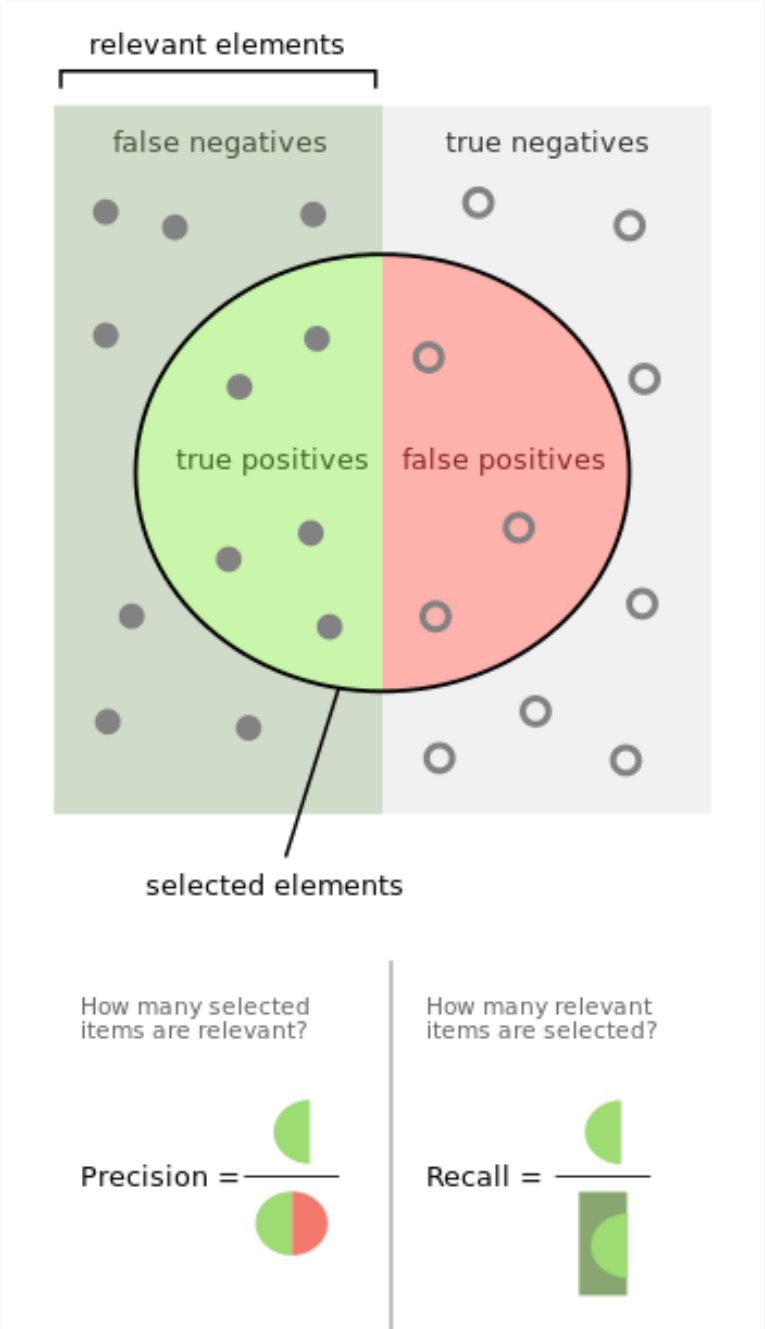
Accuracy: az összes predikció közül hány esetben értett egyet a program a kiértékelő adatbázisban szereplő címkével?

Egyszerű program: macskafelismerő

- TP: hit
  macskát mutattunk, macskát mondott
- FN: type II errror, miss, underestimation
  macskát mutattunk, nem mondott semmit
- FP: type I error, false alarm, overestimation
  kutyát mutattunk, macskát mondott
- TN: correct rejection
  kutyát mutattunk, nem mondott semmit

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

A „jóság" mérőszámai



A „jóság" mérőszámai

Hogy torzíthatjuk az eredményeket?

- a **pontosság** (precision, ami azt mutatja, hogy a találatokból hány volt eredetileg jó) növelésére: a program sose mond semmit
- a **fedés** (recall, ami azt mutatja, hogy az eredetileg jók közül hányat találtunk meg) növelésére: a program mindig macskát mond

Ellenszer: **F-mérték** (F-measure, F-score): a pontosság és a fedés harmonikus közepe, átlaga

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

Végre megérkeztünk: a korpusz

Mi az a korpusz?

A korpusz ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. A szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nemcsak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat). Az MNSZ a mai magyar írott köznyelv általános célú reprezentatív korpusza kíván lenni. Az MNSZ lényegi tulajdonsága, hogy minden szó mellett feltünteti a szótövet, a szófajt és a szó morfológiai elemzését is. A szótő, szófaj és elemzés megállapítása és az elemzések egyértelműsítése automatikus gépi eszközökkel történik. A rendszer megbízhatósága kb. 97,5%-os, így az összes szóalak kb. 2,5%-a hibásan van elemezve. Ennél jobb eredményt csak a kézi elemzés biztosíthatna, ami ekkora méretű anyag esetén megvalósíthatatlan.

(`http://corpus.nytud.hu/mnsz/bevezeto_hun.html`)

Kulcsszavak

- reprezentatív
- elektronikus formában tárolt
- nyelvészeti szempontból hasznos

Reprezentativitás

- a korpusz nem szövegek véletlen halmaza, hanem tudatosan megtervezett gyűjtemény
- a reprezentativitás megvalósítható?
- pl. egy általános nyelvi korpusz esetén olyan arányban tartalmazzon mindenféle szöveget, amilyen arányban a nyelvhasználatban is előfordulnak (diákszlengtől kezdve a filozófiai értekezéseken át a mikrohullámú sütő használati utasításáig)
- egyetlen nyelvre nézve sem áll rendelkezésünkre pontos statisztika, így a korpuszokat alkotó alkorpuszok százalékos aránya teljességgel önkényes
- kiegyensúlyozott korpusz

Mintavétel

- a kutatás tárgya határozza meg a korpusz összetételét
- minél jobban körülhatárolható a kutatási kérdés, annál könnyebb döntéseket hozni a korpusz tartalmáról

Méret

- méret általában egyenlő a szószámmal
- type és token: típus és példány
- hapax legomenon (pl. hapax legomena): egy szövegben csak egyszer előforduló szó
- szóalakok helyett lemmák

A korpuszok fajtái

A mintavétel módja szerint:

- statikus (Brown, LOB)
- dinamikus (COBUILD 1980 óta, az első korpuszalapú szótár)
- monitor

A felhasználás módja szerint:

- általános: egy nyelv minél hitelesebb reprezentálása, elsősorban a lexikográfusoknak (MNSZ, British National Corpus (BNC))
- speciális (Hong Kong Corpus of Conversational English (HKCCE))
- összehasonlítható (comparable) (Brown korpusz és a „Brown-klónok")
- párhuzamos (parallel) (`https://hunglish.hu/`)
- történeti vagy diakrón (Magyar Történeti Korpusz (`http://clara.nytud.hu/mtsz/`), ómagyar korpusz (`http://omagyarkorpusz.nytud.hu/`))

Jogi kérdések

- a szerzői jog tulajdonosának előzetes írásbeli beleegyezése nélkül jogellenes mind fénymásolatot, mind pedig elektronikus másolatot készíteni. Manapság ez nem csak teljes művekre, cikkekre, hanem részletekre is vonatkozik.
- EU: az írásművek a szerző halála után 70 évvel válnak szabadon felhasználhatóvá. Ezt megelőzően az írásmű felhasználásához a jogtulajdonos engedélye szükséges.
- a szövegek hasznosíthatóságával kapcsolatban a licenc ad tájékoztatást

Átírás és annotáció

- a beszéd átírása
- annotáció: minden olyan információ és jel, amelyet az eredeti szöveg (akár írott akár beszélt nyelvi) nem tartalmazott, de a korpusz készítésekor vagy feldolgozása során a szövegbe bekerült
- leggyakrabban: szófaji címkék (POS-tag) és parsing, de egyébként bármi (NER, szentiment, ortografikus, fonetikai, prozódiai, szemantikai, diskurzus, pragmatikai, stilisztikai annotáció)
- gold standard korpusz: kézi annotáció
- silver standard korpusz: nem ellenőrzött gépi annotáció

Min múlik a minőség?

- a kézzel annotált korpuszokat tanító- vagy kiértékelőanyagként használják felügyelt gépi tanulással működő eszközök számára
- felügyelt gépi tanulással működő rendszerek sikeressége a tanítóanyag minőségén múlik
- csak olyan feladatokat lehet felügyelt gépi tanulással megoldani, amelyeket az ember is képes elvégezni
- csak olyan nyelvi jelenségekhez tudunk kézi annotációt készíteni, amelyeket eléggé megértettünk ahhoz, hogy pontosan le tudjuk írni őket
- megbízható az annotáció, ha a jelenségek leírását több annotátor is hasonlóképpen megértette és ez alapján hasonlóképpen kódolják az egyes jelenségeket
- a feladatleírásnak tehát érthetőnek kell lennie az annotátorok számára, akik ideális esetben egyetértenek az egyes jelenségek címkézésében

Az annotátorok közti egyetértés

- a cél a minél magasabb annotátorok közötti egyetértés
- minél egyszerűbben leírható nyelvi jelenség annotálásáról van szó, annál könnyebb magas annotátorok közötti egyetértést elérni, a nyelvi jelenség összetettségével az egyetértés mértéke is könnyen csökken
- mitől lehet alacsony?
  - a feladat megfogalmazása nem egyértelmű vagy nem teljes
  - az annotátoroknak túl sok kategóriát kell kezelniük
  - átláthatatlan felületen kell dolgozniuk

Az annotátorok közti egyetértés

- az annotátorok (vagy kódolók), amikor kategóriákat rendelnek egyes elemekhez, szubjektív döntéseket hoznak
- ha az annotátorok egyetértenek az egyes elemekhez rendelt kategóriákban, akkor az adat megbízható, és ha a kódolók következetesen hasonló eredményt produkálnak, akkor hasonlóképpen értették meg a feladatot és az annotálási útmutatót, ezért a továbbiakban is hasonló eredményeket várhatunk tőlük
- *megfigyelt egyetértés*: azt mutatja meg, hogy az esetek hány százalékában értett egyet a két kódoló
- DE! nem elég, ha két kódoló egyetért, hiszen mindketten tévedhetnek is
- a címkék számának csökkentésével növekszik a megfigyelt egyetértés, ráadásul nem érzékeny az egyes címkék eltérő gyakoriságára
- megoldás: valószínűség-korrigált együtthatók, amelyek számolnak a véletlen eseményekkel is

Az annotátorok közötti egyetértés: koefficiensek

Különböző mérőszámok az egyetértésre:

- megfigyelt egyetértés
- S (Bennett, Alpert és Goldstein 1954): minden kategória ugyanolyan valószínű, a kategóriák között egyenletes eloszlást feltételez
- π (Scott, 1955): kategóriánként eltérő, de kódolók között megegyező eloszlás
- κ (Cohen, 1960): kategóriánként és kódolónként eltérő eloszlás, ez már kezeli az elfogultságot
- α (Krippendorff, 1980): nem csak az egyetértést vizsgálja, hanem az egyet nem értés különböző fokozatait

Landis-Koch skála (1977)

- -0.0: poor
- 0.0-0.2: slight
- 0.2-0.4: fair
- 0.4-0.6: moderate
- 0.6-0.8: substantial
- 0.8-1.0: almost perfect

NerKor

NER

tulajdonnév-felismerés (named entity recognition, NER):

- információkinyerés: a számítógépes nyelvészet egyik fontos alterülete; célja, hogy strukturálatlan szövegből automatikusan hozzájussunk a számunkra értékes információhoz
- egy bemeneti tokensorozatban kell megnevezett entitást (named entity, NE) alkotó intervallumokat kijelölünk, ezeket véges sok kategóriába besorolva
- egy gépi tanuló algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon
- ezért van szükség nagy méretű kézzel annotált korpuszokra

NerKor

- 1 millió token
- gold standard (kézzel annotált)
- NER annotáció, az 1/5-én morfológiai egyértelműsítés is
- szövegek vegyes doménekből
- szabadon felhasználható (CC-BY-SA 4.0. licenc)
- `https://github.com/nytud/NYTK-NerKor`

Eddigi gold standard NER korpuszok, méretek

- SzegedNER: gazdasági rövidhírek, kb. 225 000 token
- Criminal NE Korpusz, gazdasági bűncselekményekről szóló HVG-cikkek, kb. 560 000 token
- NerKor: szövegek 5 műfajból (egyenletes szövegválogatás), 1 millió token

A NerKor alkorpuszainak mérete

műfaj	morph/no-morph	fájl	mondat	token
fiction	morph	0	0	0
	no-morph	122	24535	203216
	össz	122	24535	203216
legal	morph	0	0	0
	no-morph	39	7632	202195
	össz	39	7632	202195
news	morph	35	477	9178
	no-morph	47	9280	204478
	össz	82	9757	213656
web	morph	398	10886	188250
	no-morph	0	0	0
	össz	398	10886	188250
wikipedia	morph	85	1618	26764
	no-morph	72	13096	194033
	össz	157	14714	220797
altogether	morph	518	12981	224192
	no-morph	280	54543	803922
	össz	798	67524	1028114

Források: fiction

- két fő szövegítpus:
  - szépirodalom (regények): MEK (`https://mek.oszk.hu/`), Project Gutenberg (`https://www.gutenberg.org/`)
  - filmfeliratok (OpenSubtitles, `opensubtitles.org`)
- nincs morfológiai elemzés

forrás	szerző-cím	tokenszám
MEK	Katherine Cecil Thurston: Chilcote képviselő	32581
MEK	Kaffka Margit: Képzelet-királyfiak	3813
MEK	Z. Tábori Piroska: Tökmag bandája	16355
P. Gutenberg	Karinthy Frigyes: Tanár Úr kérem	23101
OpenSubtitles	OpenSubtitles	127366
ÖSSZESEN		203216

Források: legal

- csak EU-s források, mert azok szabadon elérhetők
- nincs morfológiai elemzés

forrás	szerző-cím	tokenszám
Opus corpus	EU Constitution	17976
Opus corpus	European Economic and Social Committee	65835
ec.europa.eu	DGT-Acquis	61658
ec.europa.eu	JRC-Acquis	56726
ÖSSZESEN		202195

Források: news

morph?	forrás	szerző-cím	tokenszám
morph	KorKorpusz/Global Voices	Global Voices	9178
no-morph	Global Voices	Global Voices	76042
no-morph	NewsCrawl Corpus	NewsCrawl Corpus	52803
no-morph	Opus corpus	Press Release Database of European Commission	75633
no-morph ÖSSZESEN			204478
ÖSSZESEN			213656

Források: Wikipédia

- KorKor korpusz morfológiailag annotált szócikkei
- HunNERwiki korpusz (silver standard)

morph?	forrás	szerző-cím	tokenszám
morph	KorKorpusz/Wikipedia	Wikipedia	26764
no-morph	hunNERwiki	hunNERwiki	194033
ÖSSZESEN			220797

Források: web

- Hungarian Webcorpus 2.0
- az e-magyarral elemzett, tehát tartalmaz morfológiai annotációt (silver standard)
- az annotációt kézzel ellenőriztük, így lett gold standard

morph?	forrás	szerző-cím	tokenszám
morph	Webcorpus 2.0	Webcorpus 2.0	188250

Fájlformátum: CoNLL-U Plus

- `https://universaldependencies.org/ext-format.html`
- `tsv` formátum, amiben egy sorban egy token szerepel, a mondathatárt üres sor jelöli, a tabbal elválasztott oszlopokban pedig a fent meghatározott egyes annotációtípusok szerepelnek
- ha egy annotációtípus az adott cellában kitöltetlen marad, akkor alulvonás (‘_’) jelöli
- nullánál több, bárhány oszlopa lehet
- a fájl első sorában jelölni kell, hogy milyen oszlopai vannak, például: `# global.columns = ID FORM LEMMA UPOS XPOS FEATS CONLL:NER`
- fájlnevek kiterjesztése: `.conllup`

A conllup oszlopai

Az oszlopaink: `FORM LEMMA UPOS XPOS FEATS CONLL:NER`, ahol

`FORM`	maga a token
`LEMMA`	a token lemmája
`UPOS`	a UD jelölési formalizmusa szerinti szófajkód
`XPOS`	az `emMorph` által kiadott morfológiai kód, amely tartalmazza a szófajkódot és a morfoszintaktikai információkat is
`FEATS`	a UD jelölési formalizmusa szerinti morfoszintaktikai jegyek
`CONLL:NER`	a NE annotáció

A NE annotáció

- 4 névkategória (`PER`, `ORG`, `LOC`, `MISC`)
- 2002-es és 2003-as CoNLL shared taskok sztenderd címkekészlete
- a címkék formátuma: `IOB2` (a CoNLL2002 annotációs formátuma)
- a formátum és névkategóriák: nemzetközileg ismertek és széles körben használtak

IOB2:

- minden név első eleme '`B-`' prefixet kap
- minden név minden nem első eleme '`I-`' prefixet kap
- a nem neveket '`O`'-val jelöljük

A NE annotálás alapelvei

- Csak tulajdonneveket annotálunk. Nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem tulajdonnévvel. Például a *József Attila Gimnázium* annotálandó, de a szövegben szereplő az *a suli* frázis nem, hiába derül ki a szövegből, hogy melyik iskolára utal.
- A nevek nem kompozicionálisak. Mivel a nevek jelölete nem a név részeinek a jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy földrajzi névként jelölendő, hiába van benne egy személynév. Ebből az következik, hogy mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.
- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.

A NE annotálás alapelvei

- A *tag-for-meaning* elvét követjük. Vagyis egy nevet mindig az aktuális kontextusnak megfelelő referenciája alapján annotálunk.
- Ha az azonosított név ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.
- A nevek képzett alakjait nem jelöljük. Nem annotálandók tehát az olyanok, mint *magyarországi, fideszes, petőfieskedő*.
- Ha a név összetétel előtagja, és az összetétel alaptagja köznév, például *Horn-kormány, Tilos Rádió-hallgatók, TA-vezérigazgató*, akkor nem annotálandók névként.
- A névhez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, például *The Hague, The Times*.
- A név rövidítése (akronim, mozaikszó, monogram) is névként annotálandó.

A NE-k sűrűsége az alkorpuszokban

műfaj	morph?	PER	LOC	ORG	MISC	NE	NE density
fiction	morph	0	0	0	0	0	
	no-morph	5224	1042	217	287	6770	0,0333
	össz	5224	1042	217	287	6770	0,0333
legal	morph	0	0	0	0	0	
	no-morph	255	1302	6840	1871	10268	0,0507
	össz	255	1302	6840	1871	10268	0,0507
news	morph	220	168	183	63	634	0,0690
	no-morph	4368	2161	5111	3636	15276	0,0747
	össz	4588	2329	5294	3699	15910	0,0744
web	morph	2826	1343	1788	2434	8391	0,0445
	no-morph	0	0	0	0	0	
	össz	2826	1343	1788	2434	8391	0,0445
wikipedia	morph	571	400	203	324	1498	0,0559
	no-morph	8321	8714	5159	3929	26123	0,1346
	össz	8892	9114	5362	4253	27621	0,1250
altogether	morph	3617	1911	2174	2821	10523	0,0469
	no-morph	18168	13219	17327	9723	58437	0,0726
	össz	21785	15130	19501	12544	68960	0,0670

NE annotálás: preproceszálás

- szépirodalom: címlapok, idegen nyelvű részek kiszűrése
- kiinduló formátum $\rightarrow$ sima szöveg
- deduplikálás
- kezelhető méretű fájlokra bontás
- elemzés az e-magyarral (tokenizálás, morfológiai elemzés, egyértelműsítés)
- NE-előcímkézés különböző NE-elemzőkkel
- konverzió az annotációs segédeszköz számára megfelelő formátumra (CoNLL2002)

NE annotálás

- feladat: a szöveghez automatikusan rendelt annotáció ellenőrzése és javítása
- a javítás során a helyes címkéket kattintással jóvá kellett hagyniuk, ha pedig a címke nem volt helyes, lehetőség volt a címke módosítására, törlésére vagy a szekvencia határainak eltolására
- 11 annotátor, akiket egy tesztfeladat elvégzése után válogattunk ki
- minden szöveget két annotátor ellenőriz/javít
- folyamatos kapcsolattartás virtuálisan (email, zoom)
- annotációs útmutató
- az annotátorpárok nem állandók az egyenletes minőség miatt
- átlagos haladási tempó: kb. 6200 token/nap
- tisztázás: a kétféle annotálás ellenőrzése, felülbírálása

NE annotálás: posztprocesszálás

- exportálás (CoNLL2002-es formátumban)
- tokenizálási hibák kézi javítása
- sanity check (minden oszlop a helyén, minden címkeformátum stimmel)
- konverzió a végformátumra (CoNLL-U Plus)

A morfológiai annotáció

- kétféle címkekészlet: emMorph és UDv2
- az előannotálás és a kézi javítás az emMorph címkekészletével történt
- az UD címkéket automatikus konverzióval állítottuk elő

emMorph: teljes morfológiai elemzést ad ki, ami magában foglalja a lemmát, a szófajkódot és a morfoszintaktikai információkat is, például:
`adtad ad [/V][Pst.Def.2Sg]`

UD: A morfoszintaktikai információkat linearizált jegy–érték párok alkotják, például:
`VERB`
`Definite=Def|Mood=Ind|Number=Sing|Person=2`
`|Tense=Past|VerbForm=Fin |Voice=Act`

A morfológiai annotálás: preprocesszálás és annotálás

- szövegválogatás: KorKorpusz (már kézzel ellenőrzött, de egy része csak egy annotátorral) + Hungarian Webcorpus 2.0 egy része (már tartalmaz elemzést, de nem ellenőrzött)
- konvertálás az annotáló felület számára megfelelő formátumra
- annotáló felület: Google Spreadsheets
- az annotátor feladata: minden token esetében döntse el, hogy az `e-magyar` `emTag` modulja a helyes szófajcímkét és tövet párosította-e hozzá. Ha nem, akkor az `emMorph` által felkínált összes alternatív tő+címke kombinációból választhat, ha pedig ezek között sem volt a kontextusnak megfelelő, akkor kézzel adja meg a helyes tövet és/vagy címkét. Emellett a tokenizálást is ellenőrizze/javítsa
- 6 annotátor (nyelvész)
- átlagos haladási tempó: 1074 token/nap

KorKor

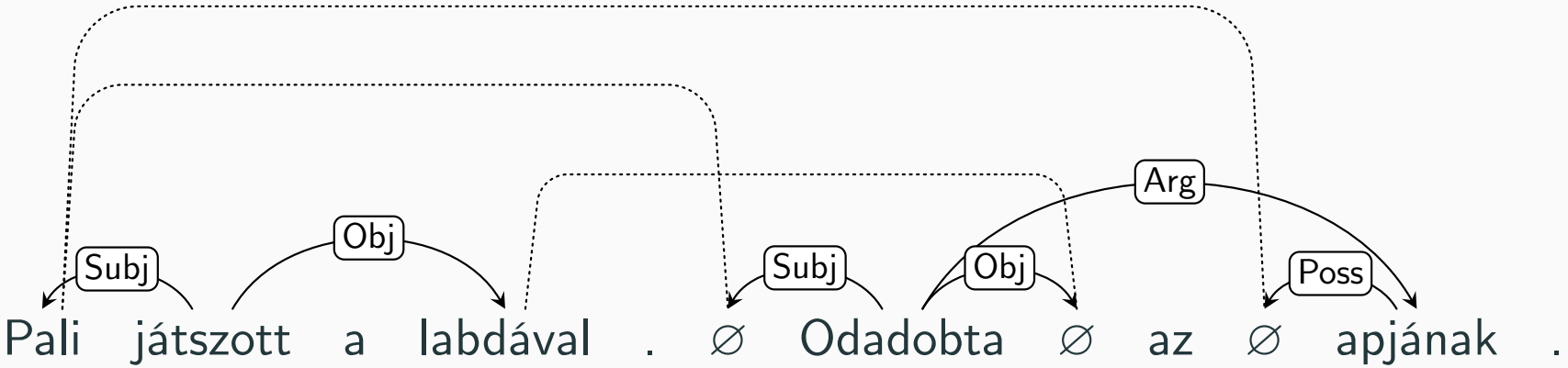
A KorKor korpusz

- **többrétegű** a szokásos elemzési rétegek (tokenizálás, tövesítés, morfológiai elemzés, egyértelműsítés, szintaktikai elemzés) mellett anafora- és koreferenciaannotációt is tartalmaz
- **kézzel annotált** minden elemzési réteg kézzel ellenőrzött minőségű
- az elemzésekhez az e-magyar legújabb változatát, az emtsv-t, valamint saját szkripteket használtunk
- **szabadon elérhető** CC-BY-4.0 licensz alatt, így bárki továbbfejlesztheti és publikálhatja az eredményeit
- a teljes munkafolyamat **reprodukálható**
- 95 dokumentum, 1 436 mondat, 31 492 token
- `https://github.com/vadno/korkor_pilot`

Koreferencia és anafora

- információkinyerés
- összekapcsolódó elemek megjelölése a szövegben
- mindkét kapcsolattípus feloldása feltétele a szöveg pontos interpretációjának
- koreferencia
  - azonos referenciájú elemek
  - szimmetrikus és tranzitív
  - fajtái: ismétlés, szinonima, holonima, meronima, hiper- és hiponima, holonima
- anafora
  - visszautaló elemek
  - kontextusfüggő
  - fajtái: személyes, mutató, vonatkozó, reflexív, reciprok stb. névmási

Zérónévmások



Előzmény: SzegedKoref

- a Szeged Korpusz egy részén készült: újsághírek és iskolai fogalmazások
- minden rétege kézzel ellenőrzött minőségű
- MSD morfológiai kódkészlet (KorKor: emMorph)
- összetevős mondatelemzés (KorKor: függőségi)
- ebben is vannak zérónévmások, de zéró igék nincsenek
- engedélykérés után kutatási célra felhasználható
- kb. 55 000 token, 4 000 mondat

I. szöveggyűjtés

- OPUS Corpus: Wikipédia és Global Voices hírszöveg
- minden szöveget átnéztünk helyesírási szempontból
- a szövegek hossza 5 és 27 mondat között, a mondatok hossza 3 és 71 token között van (az írásjeleket külön tokennek számolva)

II. elemzés és ellenőrzés

- elemzés az emtsv-vel: emToken, emMorph, emTag
- előfeldolgozó szkript és feltételes formázások
- kézi ellenőrzés a Google Spreadsheets-ben
- kézi ellenőrzés három nyelvész annotátorral
  - az emTag kimenetét
  - az emMorph kimenetének segítségével
- annotátorok közötti egyetértés
  - 4 315 token mindhárom annotátor által ellenőrizve
  - 0,976 (Krippendorff alfájában)
- utófeldolgozó szkript
  - tokenizálás javítását szolgáló parancsok automatikus értelmezése
  - előkészítés a következő elemzési fázisra

III. elemzés és ellenőrzés

- elemzés az emtsv-vel: emmorph2ud, emDep
- formátumátalakítás CoNLL-U formátumra az emtsv emCoNLL moduljával
- kézi ellenőrzés három nyelvész annotátorral
- WebAnno webalapú annotációs eszköz

IV. hiányzó igék beillesztése

- zéró létigék és elliptált igék beillesztése a teljes elemzésért
- új tokenként kerülnek be a tsv-be oda, ahol múlt időben testes létigeként jelennének meg
- a mondatfában plusz ágak jelennek meg
- saját, kombinált ID-t kapnak

A sorozat főhőse Papyrus $\varnothing_{van}$*, aki egy ifjú halászlegény* $\varnothing_{van}$*.*

Öccse miniszteri posztot vállalt, majd elnöki pozíciót $\varnothing_{vállalt}$*.*

V. zéró névmások beillesztése

- saját szkript illeszti be
- szabályalapú, a következő helyekre illeszt névmást:
  - finit ige alanyának, ha annak nem volt testes alanya
  - határozott ragozású finit ige tárgyának, ha annak nem volt testes tárgya
  - birtok birtokosának, ha annak nem volt testes birtokosa
  - ragozott és ragozatlan infinitívusz alanyának
- a mondatfában plusz ágak jelennek meg
- a névmások saját, kombinált ID-t kapnak
- a morfológiai jegyeiket automatikusan kapják
- a program az emtsv moduljaként is futtatható (emZero)

VI. névmási anaforák beillesztése

- saját szabályalapú szkript illeszti be
- egyelőre csak a személyes névmások előzményét keresi, a többi típust kézzel kell beilleszteni

VII. ellenőrzés és koreferencia

- a zérónévmások és az anaforák kézi ellenőrzése
- a koreferenciakapcsolatok kézi beillesztése
- négy nyelvész annotátorral
- Google Spreadsheets-ben, feltételes formázásokkal
- elő- és utófeldolgozó szkript
- a koreferenciakapcsolatokban az előzmény testes szó
- a névmások előzménye lehet tartalmas szó vagy névmás
- az anaforikus- és koreferenciakapcsolatok nem láncot képeznek a szövegen át, hanem elágazásokat, kitérőket is tartalmaznak

Címkekészlet

Anaforatípusok

- szokásosan jelölt anaforatípusok: személyes, mutató, kölcsönös, visszaható, vonatkozó és birtokos
- plusz típusok
  - általános
    *... a Kínai Kommunista Párt egyik volt vezetője, akit hazaárulás miatt **elítéltek** ...*
  - beszélő és címzett
    *A születésnapi ajándékoknak is nagyon **örülünk**, ha **szeretnéd** támogatni a munkánkat, **küldj nekünk** adományt ...*

Koreferenciatípusok

- valódi koreferencia (ismétlés, szinonima, hiper- és hiponima stb.) egy típussal jelölve
- rész-egész viszony

Nehéz esetek

*Három hónap telt el az **újságíró házaspár** meggyilkolása óta. A **holttesteket** már exhumálták is ...*

***Papyrus** bátor és megmenti **Thèti-Chèri-t**. **A két egymásra lelt barát** küldetést kap az istenektől ...*

Zéró elemek

zéró elem	előfordulás
létige	463
elliptált ige	25
alany	2 346
tárgy	260
birtokos	914

Kapcsolatok

kapcsolat	előfordulás
személyes	1 497
mutató	147
kölcsönös	11
visszaható	18
vonatkozó	447
birtokos	0
általános	316
beszélő	5
címzett	1
koreferencia	1 582
rész-egész	202

Munkaszervezés

mindhárom kézi ellenőrzési fázisban:

- munkaidő-nyilvántartás
- folyamatos kommunikáció
- frissülő annotálási útmutató és GYIK

	perc/dokumentum
emTag	0:24:13
emDep	0:29:54
koreferencia	0:23:23