

Különbozó módszerek szövegek automata feldolgozására

Szilágyi Péter - 2007

Tartalom

- ▶ Feldolgozás szavak szintjén

- ▶ Reguláris kifejezések
- ▶ Automaták és jelátalakítók
- ▶ “N-gram”-ok
- ▶ Szóosztályok

- ▶ Feldolgozás mondatok szintjén

- ▶ Szöveggörnyezetmentes nyelvtan
- ▶ Shrdlu



Reguláris kifejezések

- ▶ Talán a legbanálisabb
- ▶ Két fontos képességet lehet kihasználni:
Helyettesítés: Adott mintát egy másikra cserélni
Memorizálás: Illeszkedés egy részét memorizálni
- ▶ Könnyedén szerkeszthetőek olyan szabályok melyek által egy program egy korlátozott környezetben eleinte értelmesnek tűnő párbeszédet tud folytatni
- ▶ Hátrányai, hogy nincs se tudás, se szemantika, ami hamar bukáshoz vezet



Reguláris kifejezések - Eliza

▶ ELIZA – Weizenbaum (1966)

Parodizálja a személyközpontú pszichológiai terápiát
(Carl Rogers 1940-1950)

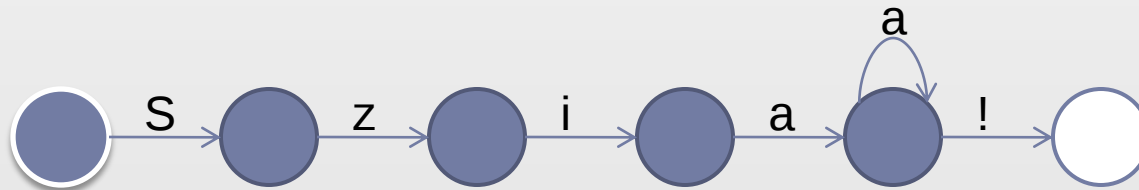
Mérföldkő mivel az első próbálkozás ilyen téren

User:	<i>Men are all alike.</i>
ELIZA:	IN WHAT WAY
User:	<i>They're always bugging us about something or other.</i>
ELIZA:	CAN YOU THINK OF A SPECIFIC EXAMPLE
User:	<i>Well, my boyfriend made me come here.</i>
ELIZA:	YOUR BOYFRIEND MADE YOU COME HERE
User:	<i>He says I'm depressed much of the time.</i>
ELIZA:	I AM SORRY TO HEAR YOU ARE DEPRESSED.



Automaták általánosan

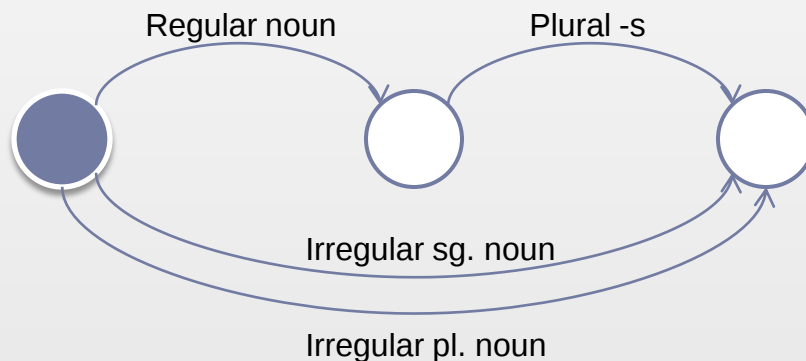
- ▶ Állapotok véges halmaza és ezek közti átmenetek
- ▶ Létezik egy jól meghatározott kezdeti állapot és egy vagy több végső
- ▶ Főként mintaillesztésre alkalmas
- ▶ Legkönnyebben gráfokkal lehet ábrázolni



- ▶ Az átmenetek nem feltétlenül korlátozódnak betűkre
 - ▶ Bonyolultabbak: nem determinisztikus, ε -átmenetek
-

Automaták szövegfeldolgozásban

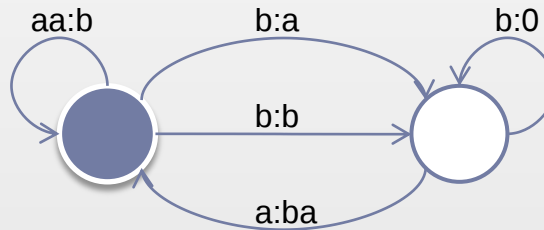
- ▶ Modellezhető egy nyelv “teljes lexikona” optimálisan
- ▶ Ábrázolhatóak bizonyos nyelvi szabályok



- ▶ Felhasználásuk mint mintaillesztés elég korlátos
- ▶ Nem elég, hogy létezik-e egy szó vagy sem, minden információra szükség van

Jelátalakítók általánosan

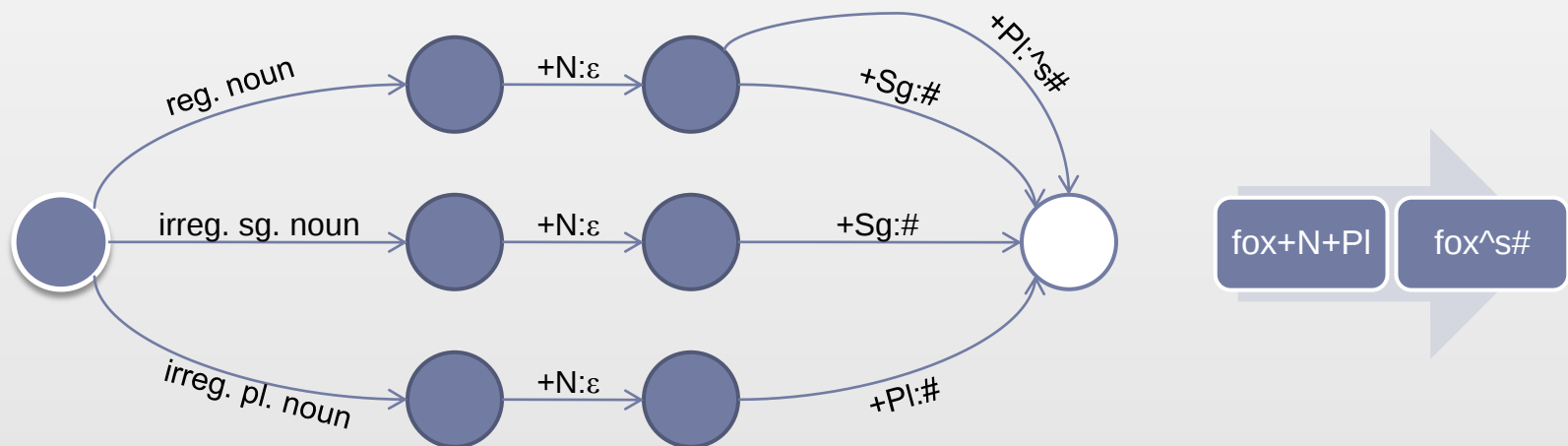
- ▶ Két különböző ábrázolást kapcsol össze, úgymond megfeleltet két szimbólumot egymásnak



- ▶ Több felhasználási köre van, például:
 - ▶ Felismerő: Sima automata, csak párokkal dolgozik
 - ▶ Fordító: Beolvas egy karakterláncot és kiír egy mást

Jelátalakítók szövegfeldolgozásban

- ▶ Jelen esetben fordító (bemenet a szöveg, kimenet a morfémákra való felbontás)
- ▶ Definiált az inverz (értelmező $\leftrightarrow$ generátor)



- ▶ Hátrány, hogy szükséges a teljes lexikon
- ▶ Javított változat: “Porter Stemmer” algoritmus (1980)

“N-gram”-ok általánosan

- ▶ Probabilisztikus modell, megpróbálja megsejteni a következő szót az előző N-1 alapján
- ▶ Alapvető eszköz zajos csatornánál vagy félreérthető szerkezetek helyes felismerésben
- ▶ A valószínűségeket szó/hang gyűjtemények alapján számolják ki (pl: szövegre a Brown corpus ill. hangra a Switchboard)
- ▶ Önmagában gyenge eredményekhez vezet, ha nem megfelelő, vagy bizonyos téren hiányos a “képzés”
- ▶ Javítható: simítások, interpolálások



“N-gram”-ok érdekesebb felhasználása

- ▶ Random generált szövegek Shakespeare összes műve alapján 😊

N	Generált szöveg
1	• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have • Every enter now severally so, let • Hill he late speaks; or! a more to leg less first you enter • Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like
2	• What means, sir. I confess she? then all sorts, he is trim, captain. • Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. • What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman? • Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
3	• Sweet prince, Falstaff shall die. Harry of Monmouth's grave. • This shall forbid it should be branded, if renown made it empty. • Indeed the duke; and had a very good friend. • Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
4	• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; • Will you not tell me who I am? • It cannot be but so. • Indeed the short and the long. Marry, 'tis a noble Lepidus.

Szóosztályok

- ▶ Dionysius Thrax of Alexandria (i.e. 100) – Tekhnê Grammatikê
- ▶ Főnév, ige, névmás, prepozíció, határozószó, névelő, kötőszó, melléknévi igenév
- ▶ Rengeteg információ a szóról is és a környezetéről is
- ▶ Több módszer az osztályok meghatározására:
 - ▶ Kézi: szabály rendszer felállítása
 - ▶ Probabilisztikus: Rejtett Markov, Maximális Entrópia
 - ▶ Más: transzformáció alapú



Szabályrendszeren alapuló osztályozás

- ▶ Szótár alapján hozzárendeli a szavakhoz az összes lehetséges “osztályukat”
- ▶ Kézi szabályrendszer alapján leszűkíti egyértelművé

Pavlov	PAVLOV N NOM SG PROPER
had	HAVE V PAST VFIN SVO
	HAVE PCP2 SVO
shown	SHOW PCP2 SVOO SVO SV
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	N NOM SG
...	

ADVERBIAL-THAT RULE
Given input: “that”
if
 (+1 A/ADV/QUANT);
 (+2 SENT-LIM);
 (NOT -1 SVOC/A);
then eliminate non-ADV tags
else eliminate ADV tag

Statisztikai modellen alapuló osztályozás

- ▶ Szósorozathoz (pl. mondat) kell hozzárendelni egy osztály sort (miden szóhoz, hogy milyen osztályba tartozik (főnév, ige, stb))
- ▶ Az összes lehetséges osztálysorozatból kell kiválasztani azt, amelyik a legvalószínűbb

$$\hat{t}^n: P(\hat{t}^n, w^n) = \max_{t^n} P(t^n | w^n)$$

$w^n := n$ hosszú szósor, $t^n := n$ hosszú osztálysor, $\hat{t}^n :=$ optimális osztálysor

- ▶ Bayes tétele alapján:

$$\hat{t}^n: P(\hat{t}^n, w^n) = \max_{t^n} \frac{P(w^n | t^n) P(t^n)}{P(w^n)}$$



Statisztikai modellen alapuló osztályozás

- ▶ Nevező konstans tehát:

$$\hat{t}^n = \operatorname{argmax}_{t^n} P(w^n | t^n) P(t^n)$$

argmax az az argumentum,
melyre a kifejezés maximális

- ▶ Feltételezzük, hogy egy szó előfordulása csak az osztályától függ

$$P(w^n | t^n) \approx \prod_{i=1}^n P(w_i | t_i)$$

- ▶ Feltételezzük, hogy a valószínűsége, hogy egy osztály megjelenjen csak az előző osztálytól függ

$$P(t^n) \approx \prod_{i=1}^n P(t_i | t_{i-1})$$



Statisztikai modellen alapuló osztályozás

- ▶ Tehát a leegyszerűsített változat:

$$\operatorname{argmax}_{t^n} P(t^n | w^n) \approx \operatorname{argmax}_{t^n} \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

- ▶ Megfigyelhető, hogy az egyszerűsítés második tagja az egy 2-gram (használnak még 3-gram-okat is)
- ▶ Ez az forma már eléggé le van egyszerűsítve, hogy programmal lehessen dolgozni vele
- ▶ Nincs teljes mértékben elhagyva a szavak sorrendje



Statisztikai modellen alapuló osztályozás

▶ Például: Bill is expected to race tomorrow.

▶ “race” mint ige:

$$P_1 = P_c \cdot P(ige|TO) \cdot P(race|ige) \cdot P(idő\ hat. | ige)$$

▶ “race” mint főnév:

$$P_2 = P_c \cdot P(főnév|TO) \cdot P(race|főnév) \cdot P(idő\ hat. | főnév)$$

$$P(főnév|TO) = 0.00047 \quad P(race|főnév) = 0.00057 \quad P(idő\ hat. | főnév) = 0.0012$$

$$P(ige|TO) = 0.83 \quad P(race|ige) = 0.00012 \quad P(idő\ hat. | ige) = 0.0027$$

$$P_1 = P_c \cdot 0.00000027 \quad P_2 = P_c \cdot 0.000000000032$$



Transzformáción alapuló osztályozás

- ▶ Szabályok alapján működik, viszont automatikusan vezeti le ezeket
- ▶ Festő analógia: először nagyvonalakban majd egyre inkább részletez
- ▶ Előre definiált “template”-ek alapján generálja a szabályokat (tanító adat is szükséges)
- ▶ Automatikus szabálygenerálás algoritmusai:
 - ▶ Inicializálunk mindent a legvalószínűbb osztállyal
 - ▶ Amíg egy bizonyos korlátot el nem érünk:
 - ▶ Geráljunk a lehetséges “template”-eket
 - ▶ Ezek alapján kiválasztjuk a legjobb transzformációt
 - ▶ Alkalmazzuk és a szabály sorozat végére tesszük
- ▶ Új adatra egyszerűen lefuttatjuk a szabály sorozatot



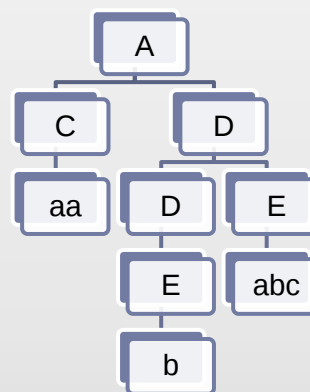
Feldolgozás mondatok szintjén

- ▶ Előzőekből láttuk, hogy fontos:
 - ▶ Szavak mondattani osztálya
 - ▶ Szavak és osztályok sorrendje
- ▶ Eddigi megközelítéssel nem lehet modellezni:
 - ▶ Alkotórészek (szócsoport mint egység)
 - ▶ Nyelvtani kapcsolatok (“mondattani összefüggések”)
 - ▶ Függőségi kapcsolatok (szavak és mondatok között)
- ▶ Szövegkörnyezetmentes nyelvtanok – kifinomult összefüggések ábrázolhatóak de mégis kezelhetőek számítógépekkel



Szövegekörnyezetmentes nyelvtan

- ▶ Wilhelm Wundt (1900)
- ▶ Chomsky (1956) illetve Backus (1959) formalizálta
- ▶ Adott egy szabály halmaz (kifejezik, hogy a nyelv szimbólumait hogyan lehet csoportosítani) és egy szó illetve szimbólum lexikon

$$A \rightarrow B \mid C + D$$
$$D \rightarrow E \mid D + E$$
$$C \rightarrow a \mid aa$$
$$E \rightarrow b \mid ab \mid abc$$


- ▶ Formális nyelv – adott gyökérből származtatható “mondatok” halmaza

Formalizált angol nyelvre példa

► Gyökérnek tekinthetjük a mondatot (S)

Szabályok	
• S	→ NP VP
• NP	→ Pronoun → Proper-Noun → Det Nominal
• Nominal	→ Nominal Noun → Noun
• VP	→ Verb → Verb NP → Verb NP PP → Verb PP
• PP	→ Preposition NP

Lexikon	
• Noun	→ flights breeze trip morning ...
• Verb	→ is prefer like need want fly ...
• Adjective	→ cheapest non-stop first latest other direct ...
• Pronoun	→ me I you it ...
• Proper-Noun	→ Alaska Baltimore Los Angeles Chicago United American ...
• Determiner	→ the a an this these that ...
• Preposition	→ from to on near ...
• Conjunction	→ and or but ...

Shrdlu

- ▶ Terry Winograd (MIT) 1968 – 1970
- ▶ Egy “kocka világ” irányítása angol szöveggel
- ▶ MacLisp-ben írták
- ▶ Megközelítőleg 250 angol szót ismer

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.



Referenciák

- ▶ Daniel Jurafsky & James H. Martin (2007). An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition
- ▶ Terry Winograd (MIT 1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language
- ▶ Douglas R. Hofstadter (1979). Gödel, Escher, Bach: an Eternal Golden Braid



Weboldalak

- ▶ <http://en.wikipedia.org/wiki/ELIZA>
- ▶ <http://en.wikipedia.org/wiki/Stemming>
- ▶ http://en.wikipedia.org/wiki/Bayesian_inference
- ▶ <http://hci.stanford.edu/~winograd/shrdlu/>



Köszönöm a figyelmet! 😊

