

pragmatika: a nyelvészetnek azon területe, mely a nyelv használatával foglalkozik.

párhuzamos korpusz: több nyelven párhuzamosan elérhető szöveges forrás, mely segítségével különféle fordítási feladatokat tudunk ellátni

poliszémia: többértelműség (egy szónak több jelentése van)

RTN: Recursive Transition Network. Egymást hívó FSA-k hálózata: az élek címkéi közt megjelenik más FSA-k „neve”. A FSA (a reguláris nyelvek) $O(n)$ idő alatt elemezhető. Az RTN valójában egy veremautomata.

Linear precedence: mi mit előzhet meg az egész nyelvtanban

Morfológia: nyelvészeti terület. aka. alaktan. a szóalakok szerkezetével foglalkozik.

jelentés-egyértelműsítés: megállapítani a többjelentésű szavak aktuális jelentését

fordító memória: A fordítómémória (translation memory, TM) szegmenspárokat (általában mondatpárokat) tárol. Forrásnyelvi szegmensek és célnyelvi fordítások párait tárolja, némi adalékinformációval (rögzítés dátuma, fordító neve, fordítás státusza stb.) Ha később egy hasonló forrásnyelvi szegmenst kell fordítani, a TM

program megtalálja és felkínálja a szegmens korábbi fordítását.

synset: szinonima halmaz, WordNet csúcsok címkéi: a jelentések, mint szinonimahalmazok

fonetika: hangtan, a beszédhangok képzése, fizikai tulajdonságai

szemantika: jelentéstan, a jelentés

homonímia: hason(ló)nevűség, azonosalakúság”. két szó – esetleg egyéb nyelvi elem, például szókapcsolat – közötti véletlen egybeesés, amikor íráskéjük (esetleg hangalakjuk) azonos, de a jelentésük különbözik

balsarok: B-t A balsarkának nevezzük, ha van A-nak B-vel kezdődő levezetése. Algoritmusok melyek balról-jobbra működnek, és elkerülik lehetetlen szerkezetek felépítését. egy olyan elemzési fát építenek fel, amely a nyelvtan kezdő szimbólumával indul, és a mondat legszélén balra (a bal sarokban) álló szó felé terjeszkednek.

immediate dominance: the case where the two nodes in question are connected by a single branch without any intervening nodes. More formally, immediate dominance is defined:

A immediately dominates B iff (= if and only if)

a. A dominates B, and

b. there is no node C, distinct from A and B, such that A dominates C and C dominates B.

fonológia: vagy fonematika a nyelvtudománynak az az ága, amely a beszédhangoknak a nyelvi rendszerben elfoglalt helyét és szerepét vizsgálja, szemben a fonetikával, amely egy adott nyelv beszédhangjainak a leírásával foglalkozik.

kompozicionalitás elve: egy komplex kifejezés jelentése a részei jelentéseinek és az őket leíró szintaktikai szerkezetnek a függvénye

szintaxis: vagy mondattan a nyelvészetben a mondatok szerkezete

Mi a közvetlen összetevős nyelvtan és a függőségi szerkezet ? Mik az előnyei és hátrányai ? (6p)

közvetlen összetevős nyelvtanok:

- előnyük a magasabb szintű kategóriák bevezetésének lehetősége
- hátrányuk a szintaktikai viszonyok egy részének „kifejezhetetlensége”

függőségi szerkezet:

- előnyük a szintaktikai függőség kifejezésének lehetősége
- hátrányuk a magasabb szintű kategóriák kezelhetetlensége

Hasonlítsd össze a top-down (TD) és a bottom-up (BU) elemzést! (6p)

TD: nem fordulhat elő, hogy ne mondat-szimbólummal végződő elemzéssel töltsük az időt, és nincs olyan részszerkezet, melynek ne lenne helye a végső fában

BU: nem fordulhat elő, hogy a bemenettel inkonzisztens szabályokkal kellene töltenie az időt, és nincs olyan részszerkezet, mely ne lenne kompatibilis a bemenet valamely részével

Optimális megoldás: valamiféle kombináció, pl. BU-elemzés előfordulási valószínűségekkal; TD-elemzés „BU-jóslással”

Mit jelent a szótár kifordíthatósága ? Mire jó? (5p)

Többféle jelentés kezelését teszi lehetővé, a bal oldali címszavak segítségével, új találati ablak behozásával. Elektronikusan érdemes végrehajtani.

Mi a hossz alapú szövegszinkronizáció lényege ? (3p)

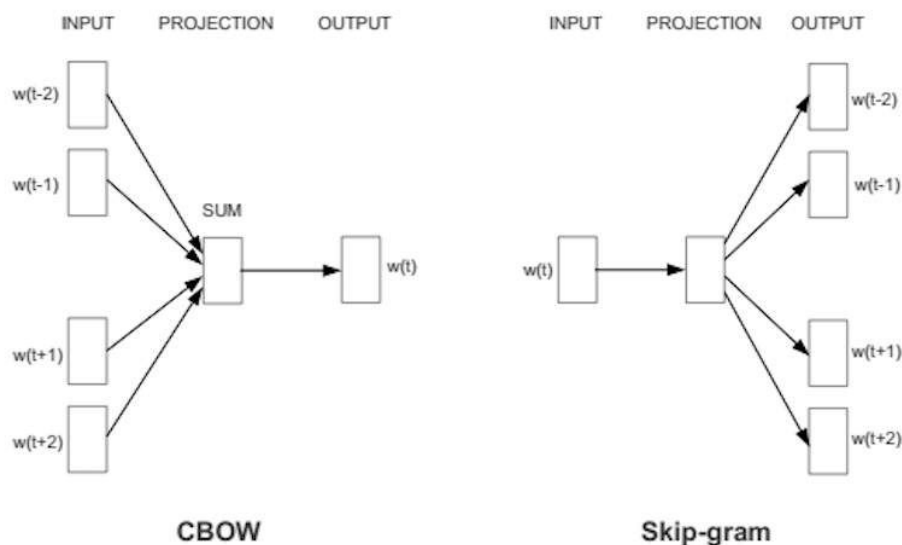
- “length-based alignment”
- csak a szövegségek hosszait vizsgálja
- valószínűségi modell
- teljes szinkronizáció
- dinamikus programozás, globálisan optimális megoldás
- robosztus

Mik azok az X-vonás nyelvtanok? Mitől jók?

- Megőrzi mind az összetevős, mind a függőségi leírás előnyeit.
- Szófajok feletti általánosítás $X = V / N / P / Adj \dots$
- Az X-nél magasabb szerkezetek is egységesen leírhatók vele
- Megkülönböztethetők különféle jellegű bővítmények

Mi az a CBOW és a Skip-gram? Mire használják őket?

- **Szóbeágyazásra használjuk:** nyelvmodellezési és jegytanulási technikák gyűjtőneve, ahol a szavakat egy n-dimenziós valós vektor reprezentálja. A leképezés neurális hálókkal, a szókörnyezet reprezentációjával történik. Mikolov (2013): word2vec egy kétrétegű neurális háló szövegrepresentációs célokra, ahol a bemenet a szövegkorpusz és a kimenet a korpusz szavait reprezentáló jegyvektorok halmaza. A word2vec maga nem mélyneurális háló, de a kimenetét mélyneurális hálók is tudják kezelni.
- (1) környezet→szó: folytonos szósák (Continuous Bag Of Words, CBOW)
- (2) szó→környezet: skip-gramm

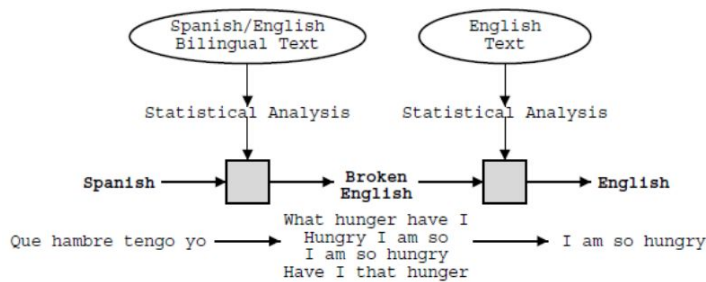


Mi a horgony alapú szövegszinkronizáció lényege?

- A horgonyok a szövegegységek egymásnak nagy pontossággal megfeleltethető pontjai (pl. nevek, számok)
- Kezdetben két horgony a szövegpár végpontjait köti össze
- Ezek után olyan szavakat keres, amelyek gyakran fordulnak elő együtt a szövegpár két oldalán nagyjából azonos pozícióban (heurisztika), majd ezekből horgonyokat határoznak meg (de pl. az egymást keresztező horgonyokat elvetik)
- A horgonyok által kisebb szinkronizálандó szegmensekre bontható a szöveg, ezeken belül újra horgonyokat keresnek
- Egyes horgonyokkal csak részleges szinkronizáció valósítható meg
- Az eredmények nem elég jók

Hogyan működik a statisztikai gépi fordítás és a szabályalapú gépi fordítás? Hasonlítsd össze őket!

statisztikai



- párhuzamos korpuszból tanul
- emberek által előre lefordított szövegek felhasználásán alapszik
- beszédtechnológiában használt Shannon-féle zajoscsatorna-modelljét vették alapul

szabályalapú:

- a felnőtt: szabályok megtanulása és szótár

- direkt fordítás
- transzfer
- interlingua



Szótárak esetén mi a perjel probléma ? Írj rá példát is!

a szótárak nem a gépnek készülnek, tudhatja-e egy program összefogni az egybetartozó kifejezést: arcára fagy/ráfagy az arcára/lefagy az arcáról a mosoly.

Hogyan működik a referenciafordítás nélküli kiértékelés ? Hasonlítsd össze a referenciafordítással történő módszerrel!

- Referenciafordítással történő kiértékelés:
 - szükséges hozzá emberi fordítás
 - drága, nem valós idejű kiértékelés
 - automatikus kiértékelő módszerek
 - nincsen szükség emberi beavatkozásra
 - nem kell ismerni a két nyelvet
 - hasonlóságot/eltérést mér
 - pl BLEU NIST METEOR etc.
- Referenciafordítás nélküli:
 - nem használ referenciát
 - valós idejű kiértékelés
 - magas korreláció az emberi kiértékeléssel
 - gépi tanuláson alapszik
 - minőségi mutatók kinyerése a forrásnyelvi és a célnyelvi mondatokból

Feladatok:

1) Unifikáció:

$hó [-elöl +DAT +INS: v]$
 $hav [-elöl -DAT -INS]$
 $füge [+elöl +INS: v]$
 $füge' [+INS: v]$

$hónak \neq fügenek$
 $hóval fügevel$ } értelmezés

$fügenek$
 $fügeval$
 $fügenak$ } elrehozható

$hó [+DAT +INS: v]$
 $hav [-elöl -DAT -INS]$
 $füge [+elöl +INS: v]$
 $füge' [+elöl +INS: v]$

$nak [+DAT -elöl]$
 $nek [+DAT +elöl]$
 $val [+INS: v -elöl]$
 $vel [+INS: v +elöl]$

$hó \rightarrow \underline{hónak}, \underline{hónek}, \underline{hóval}, \underline{hóvel}$

$hav \rightarrow \emptyset$

$füge \rightarrow \underline{fügenek}, \underline{fügevel}$

$füge' \rightarrow \underline{fügenek}, \underline{fügevel}$

	a	aa	igen nem (elő, utó) (elő, utó)
DAT	+	-	
elő	előképzett		+
INS: ?	amilyen hanggal kezdődik		-

[2] Levenshtein távolság:

level	- level	: 2	lível / nível
wood	- would	: 2	woud / would
asztal	- asatl	: 3	astal / asatl
plain	- plian	: 2	plein / plian
weather	- whether	: 2	whather / whether
blue	- blues	: 2	blue / blues

[3] Balsarok:

Terminálisok: Det, N, V

S → NP VP

NP → Det Nom

NP → Nom

Nom → N Nom

Nom → N

VP → V

VP → V NP

balsarok Det N

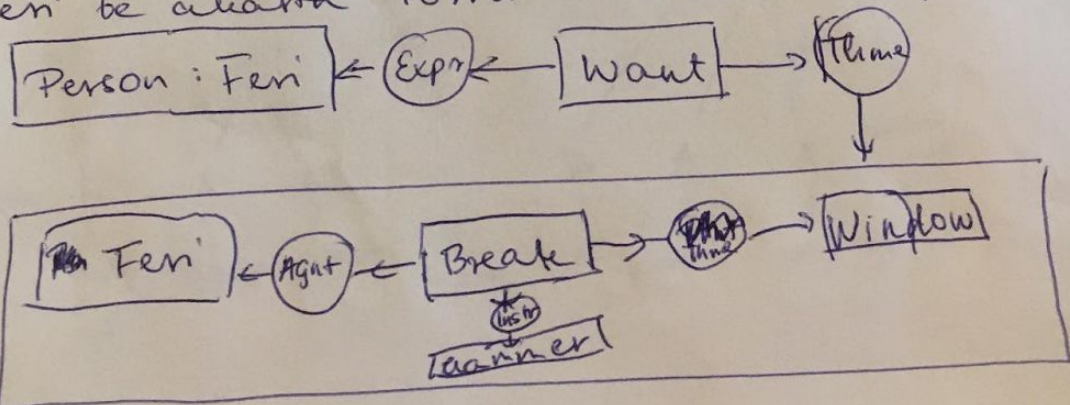
NP → Det N

NP → Det Nom

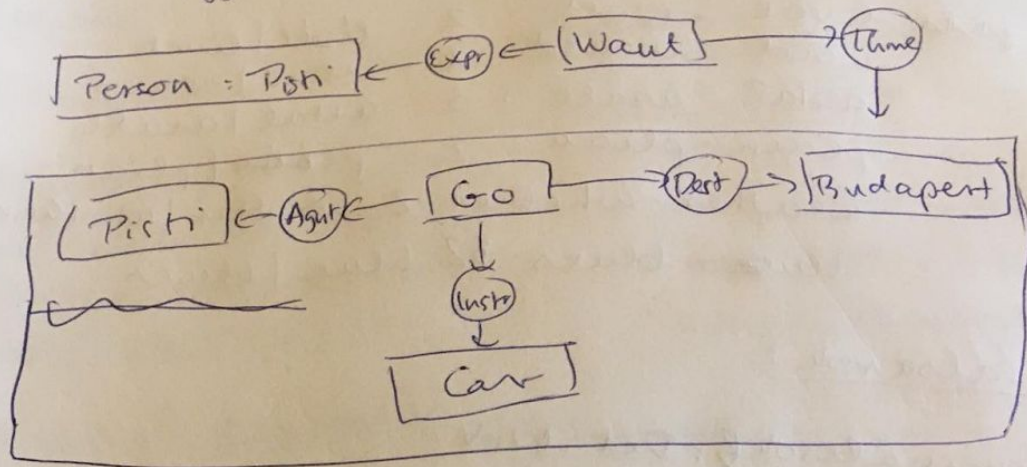
↓
N

[4] koncepciógráf:

Feri te akarta tőmi az ablakot egy kapapice



Pisti egy autóval szeretne elmenni Bp-re.



5

Word Net

szállásbázis, mosdó, kőrté, sárgabarack, kajszibarack,
gyümölcs, vilmoskőrté, lélegzők és
~~szállásbázis~~ aszpermyz,
(kőrté, sárgabarack, kajszibarack, vilmoskőrté) → gyümölcs

vilmoskőrté → kőrté → gyümölcs

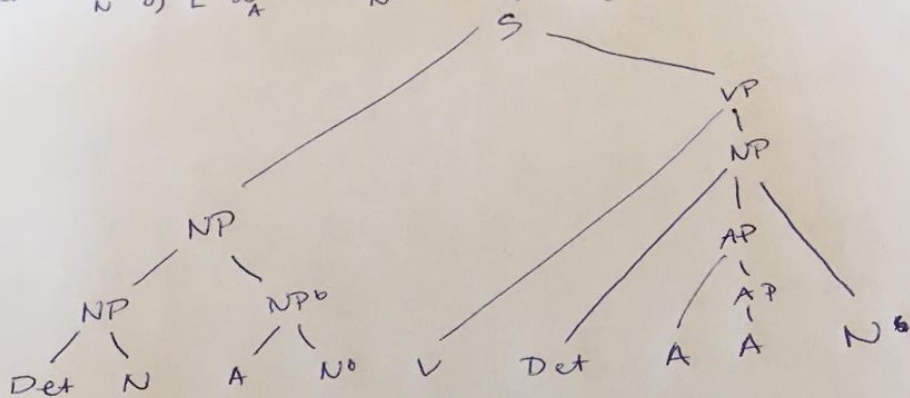
entailment: lélegzők → él

Synonym: szállásbázis → mosdó

6

Stella

[A₂ Det ontály] [legjobb A tanulóját] [focizik V] [a Det két A] [pókos] barátival N_s.



[A Det csapat] [leggyengébb A játékosát] [kicsélezte V] a Det [lassú A] [mozgási A] nézőket N.

