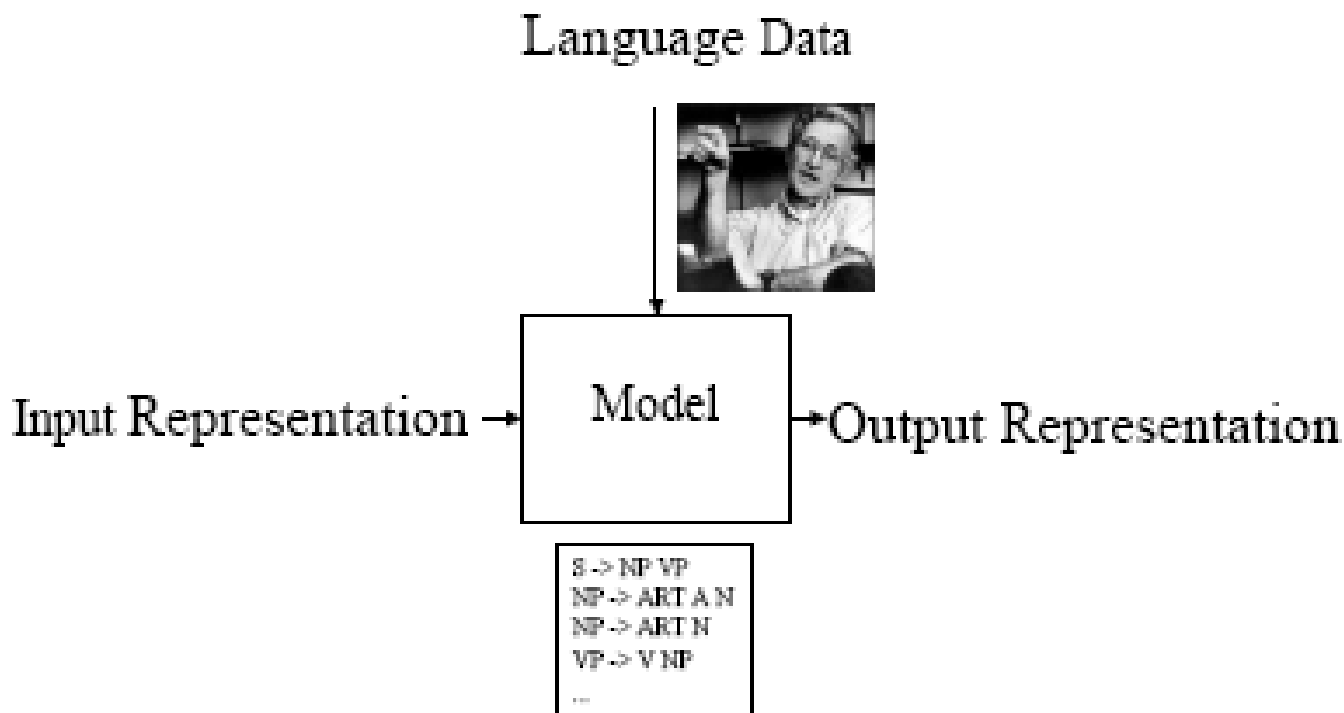


ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΣΥΝΤΑΞΗ: ΣΤΟΧΑΣΤΙΚΕΣ ΜΕΘΟΔΟΙ
STOCHASTIC PARSING

Deductive (Συναγωγή) Route

- Στην προηγούμενη διάλεξη
 - υλοποιήσαμε ένα μοντέλο κανόνων (γραμματική) και
 - χρησιμοποιήσαμε συμπερασμό με βάση το μοντέλο κανόνων (rule-based reasoning) για να αναλύσουμε καινούρια δεδομένα (προτάσεις)

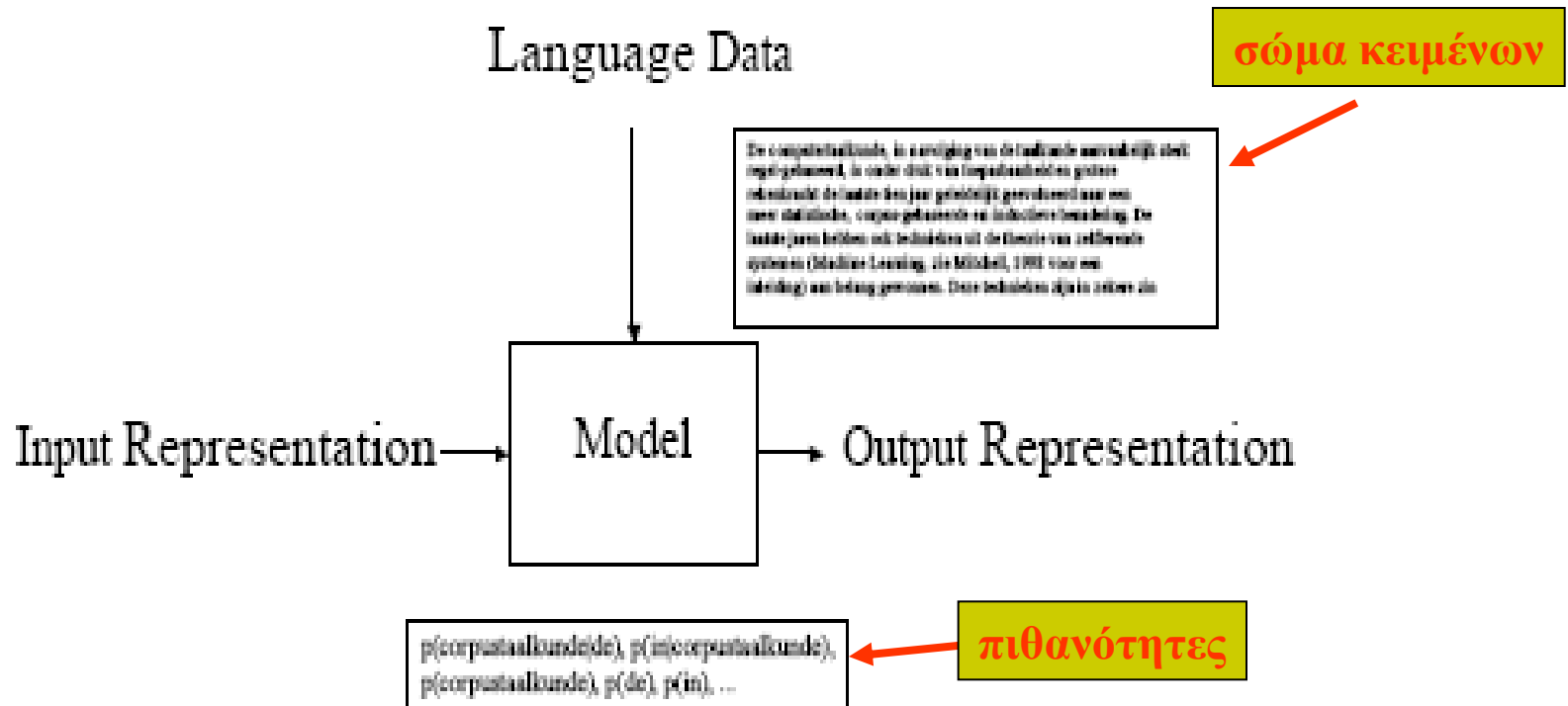


Στοχαστική Συντακτική Ανάλυση

- Μέχρι τώρα είδαμε πώς μπορούμε να κατασκευάσουμε χειρωνακτικά μία γραμματική
- Τέτοιες γραμματικές
 - Απαιτούν πολύ κόπο και χρόνο
 - Δεν μπορούν να αντεπεξέλθουν σε κείμενα του πραγματικού κόσμου
- Οι στοχαστικές μέθοδοι προσφέρουν μία καλή λύση αυτοματοποίησης.

Inductive Route (Επαγωγή)

- Στην σημερινή διάλεξη θα δημιουργήσουμε ένα στοχαστικό μοντέλο από ένα σώμα παραδειγμάτων (κειμένων)
- Με χρήση στατιστικού συμπερασμού (inference) από το μοντέλο θα επεξεργαστούμε καινούρια παραδείγματα



Θεωρία Πιθανοτήτων και Γλωσσική Τεχνολογία

- **Επίλυση ασάφειας μέρους-του-λόγου**
 - *fly*: ουσιαστικό ή ρήμα
- **Σχολιασμός κειμένου**
 - Tagging, text chunking
- **Εμπλουτισμός υπαρχόντων γραμματικών**
 - Εισαγωγή πιθανοτήτων στους κανόνες
- **Επίλυση ασάφειας έννοιας λέξεων**
 - *βιβλιοθήκη*: έπιπλο, κτίριο
- **Κατανόηση ομιλίας**
 - Εύρεση της πιο πιθανής ακολουθίας λέξεων
- **Ανάκτηση πληροφορίας**
 - Εύρεση των πιο σχετικών κειμένων με μία ερώτηση

Σώματα Κειμένων (Corpora)

- Οι στατιστικές μέθοδοι βασίζονται στην ανάλυση μεγάλων σωμάτων κειμένων που έχουν σχολιασθεί χειρονακτικά (manually annotated)
- Εξάγονται στατιστικές μετρήσεις
 - *fly*: είναι ουσιαστικό στο 95% των περιπτώσεων που προηγείται άρθρο (*the fly*)
- Αυτές οι μετρήσεις βοηθούν στην ανάλυση νέων (μη-σχολιασμένων) κειμένων
- Χρησιμοποιούμε τη θεωρία πιθανοτήτων για να βρούμε ποια είναι η πιο πιθανή λύση

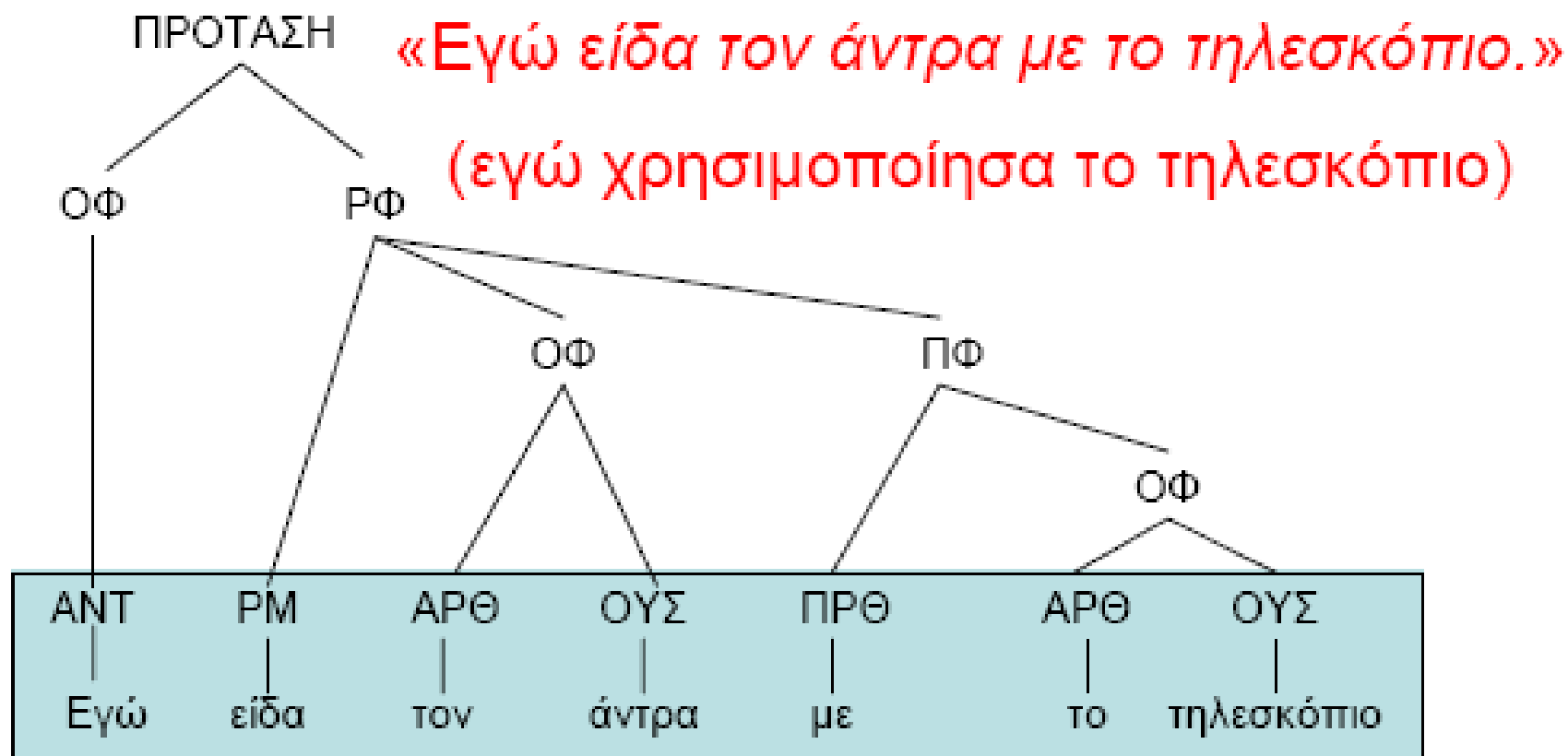
Θεωρία Πιθανοτήτων

- $0 \leq P(x) \leq 1$
- Πιθανότητα υπό συνθήκη
 - $P(A | B) = P(A \& B) / P(B)$
 - $P(A | B) = P(B | A) * P(A) / P(B)$ [Bayes rule]
- Ανεξάρτητα γεγονότα
 - $P(A | B) = P(A)$
- Αν σε ένα σώμα κειμένων έχουμε
 - 150 εμφανίσεις της λέξης *flies* ως ουσιαστικό
 - 50 εμφανίσεις της λέξης *flies* ως ρήμα
 - $P(\text{category}=\text{noun} | \text{word}=\text{flies}) = 150/200 = 0.75$

Σχολιασμός Μερών του Λόγου

- Προηγείται της συντακτικής ανάλυσης
- Σε κάθε λέξη αποδίδεται μία ετικέτα (tag) μέρους-του-λόγου
 - part-of-speech (POS) tagging
- Χρησιμοποιούμε πληροφορία από το περιβάλλον της λέξης (γειτονικές λέξεις)
 - $P(\text{cat}_n=\text{noun} \mid \text{word}_n=\textit{flies} \ \& \ \text{word}_{n-1}=\textit{the})$

Παράδειγμα: POS Tagging



Ασάφεια (Ambiguity) Μέρους-του-λόγου

- Δημιουργεί πρόβλημα κυρίως στα Αγγλικά
- Οι λέξεις μπορεί να αντιστοιχούν σε περισσότερα από ένα POS tags
 - *The back door* (επίθετο)
 - *On my back* (ουσιαστικό)
 - *Win the voters back* (επίρρημα)
 - *Promised to back the bill* (ρήμα)
- 90% ακρίβεια αν στην κάθε λέξη αποδίδεται πάντα το πιο συχνό tag

POS Tagging

- Η πρώτη εργασία επεξεργασία φυσικής γλώσσας που αντιμετωπίστηκε επιτυχώς με στατιστικές μεθόδους
- Πολλές διαφορετικές μέθοδοι έχουν εφαρμοστεί
- Η αξιολόγηση είναι εύκολη (πόσα tags αποδόθηκαν σωστά)
- Τα tags μπορεί να αποδίδουν πιο σύνθετη μορφολογική πληροφορία (πτώση, γένος) ώστε να έχει μεγαλύτερη χρησιμότητα για άλλες γλώσσες

Χρησιμότητα POS Tagging

- Απαιτεί ελάχιστο χρόνο και μπορεί να βελτιώσει την απόδοση πιο σύνθετων συστημάτων
 - Προεπεξεργασία σε ένα parser
 - Δημιουργία επιτονισμού
 - récord (ουσιαστικό), recórd (ρήμα)
 - Εύρεση του λήμματος μιας λέξης
 - saw (ουσιαστικό) -> saw
 - saw (ρήμα) -> see

Αγγλικά tagsets: Penn Treebank (45)

CC	Coord Conjunction	<i>and, but, or</i>	NN	Noun, sing. or mass	<i>dog</i>
CD	Cardinal number	<i>one, two</i>	NNS	Noun, plural	<i>dogs</i>
DT	Determiner	<i>the, some</i>	NNP	Proper noun, sing.	<i>Edinburgh</i>
EX	Existential there	<i>there</i>	NNPS	Proper noun, plural	<i>Orkneys</i>
FW	Foreign Word	<i>mon dieu</i>	PDT	Predeterminer	<i>all, both</i>
IN	Preposition	<i>of, in, by</i>	POS	Possessive ending	<i>'s</i>
JJ	Adjective	<i>big</i>	PP	Personal pronoun	<i>I, you, she</i>
JJR	Adj., comparative	<i>bigger</i>	PP\$	Possessive pronoun	<i>my, one's</i>
JJS	Adj., superlative	<i>biggest</i>	RB	Adverb	<i>quickly</i>
LS	List item marker	<i>1, One</i>	RBR	Adverb, comparative	<i>faster</i>
MD	Modal	<i>can, should</i>	RBS	Adverb, superlative	<i>fastest</i>

RP	Particle	<i>up, off</i>	WP\$	Possessive-Wh	<i>whose</i>
SYM	Symbol	<i>+, %, &</i>	WRB	Wh-adverb	<i>how, where</i>
TO	"to"	<i>to</i>	\$	Dollar sign	<i>\$</i>
UH	Interjection	<i>oh, oops</i>	#	Pound sign	<i>#</i>
VB	verb, base form	<i>eat</i>	"	Left quote	<i>' , "</i>
VBD	verb, past tense	<i>ate</i>	"	Right quote	<i>' , "</i>
VBG	verb, gerund	<i>eating</i>	(Left paren	<i>(</i>
VBN	verb, past part	<i>eaten</i>)	Right paren	<i>)</i>
VBP	Verb, non-3sg, pres	<i>eat</i>	,	Comma	<i>,</i>
VBZ	Verb, 3sg, pres	<i>eats</i>	.	Sent-final punct.	<i>. ! ?</i>
WDT	Wh-determiner	<i>which, that</i>	:	Mid-sent punct.	<i>: ; — ...</i>
WP	Wh-pronoun	<i>what, who</i>			

Αγγλικά tagsets: Brown Corpus (87)

CC: conjunction, coordinating

and or but plus & either neither nor yet 'n' and/or minus an'

CD: numeral, cardinal

two one 1 2 1913 8 five three million 87-31 29-5 1,119 fifty-three 7.5 bill 125,000

CD\$: numeral, cardinal, genitive

1960's 1961's .404's

CS: conjunction, subordinating

that as after whether before while like because if since for than altho until so unless

DO: do

DO*: don't

Παράδειγμα επισημείωσης ΜΤΛ

- Jaguar shares stood at 405 pence before Ford 's initial announcement , but the subsequent takeover frenzy has driven them up.
- Jaguar/**NN** shares/**NNS** stood/**VBD** at/**IN** 405/**CD** pence/**NN** before/**IN** Ford/**NNP** 's/**POS** initial/**JJ** announcement/**NN** ,/, but/**CC** the/**DT** subsequent/**JJ** takeover/**NN** frenzy/**NN** has/**VBZ** driven/**VBN** them/**PRP** up/**RB** ./.

Το Στατιστικό Μοντέλο

- Σε μία ακολουθία λέξεων $w_1..w_n$ θέλουμε να αποδώσουμε την πιο πιθανή ακολουθία κατηγοριών $c_1..c_n$ που μεγιστοποιεί την πιθανότητα
 - $P(c_1..c_n | w_1..w_n)$
 - $P(\textit{art noun verb} | \textit{the man talks})$
 - $P(c_1..c_n | w_1..w_n) = P(w_1..w_n | c_1..c_n) * P(c_1..c_n) / P(w_1..w_n)$
- Η $P(w_1..w_n)$ είναι ανεξάρτητη των $c_1..c_n$. Έτσι μπορούμε να απλοποιήσουμε το μοντέλο
 - $\operatorname{argmax}_{c_1..c_n} (P(w_1..w_n | c_1..c_n) * P(c_1..c_n))$

Ένα πιο απλό μοντέλο: N-grams

- Αν υποθέσουμε ότι για μία λέξη η πιθανότητα μιας κατηγορίας c_i εξαρτάται μόνο από την κατηγορία της προηγούμενης λέξης c_{i-1}
 - $P(c_1..c_n) = \prod_{i=1..n} P(c_i | c_{i-1})$
 - Μοντέλο bigram
- Αν θεωρήσουμε τις δύο προηγούμενες λέξεις έχουμε το μοντέλο trigram
- Για $n-1$ προηγούμενες λέξεις: μοντέλο n -gram
 - $P(c_1..c_n) = \prod_{i=1..n} P(c_i / c_1, \dots, c_{n-1})$

Ένα ακόμα πιο απλό μοντέλο: Naïve Bayes

- Υποθέτουμε ότι η πιθανότητα μιας λέξης είναι κατά βάση ανεξάρτητη των γειτονικών λέξεων
 - $P(w_1..w_n | c_1..c_n) = \prod_{i=1..n} P(w_i | c_i)$
 - Η πιθανότητα της φράσης “the man talks” δεδομένης της ακολουθίας “art noun verb” είναι $P(the | art) * P(man | noun) * P(talks | verb)$
- Το στατιστικό μοντέλο τώρα γίνεται
 - $\prod_{i=1..n} P(c_i | c_{i-1}) * P(w_i | c_i)$

Εύρεση Bigrams

- Υποθέτουμε ότι έχουμε ένα corpus σχολιασμένο με POS tags
- Υπολογίζουμε τα bigrams
 - $P(\text{cat}_i = \text{verb} \mid \text{cat}_{i-1} = \text{noun}) =$
 $P(\text{cat}_i = \text{verb} \ \& \ \text{cat}_{i-1} = \text{noun}) / P(\text{cat}_{i-1} = \text{noun}) =$
αριθμός ουσιαστικών που ακολουθούνται από ρήματα
προς το συνολικό αριθμό των ουσιαστικών
- Ορίζουμε $\text{cat} = 0$ για την εκκίνηση μιας πρότασης
 - $P(\text{cat}_i = \text{verb} \mid \text{cat}_{i-1} = 0) =$ αριθμός ρημάτων που ξεκινούν
προτάσεις προς συνολικό αριθμό προτάσεων

Εύρεση Λεξιλογικών Πιθανοτήτων

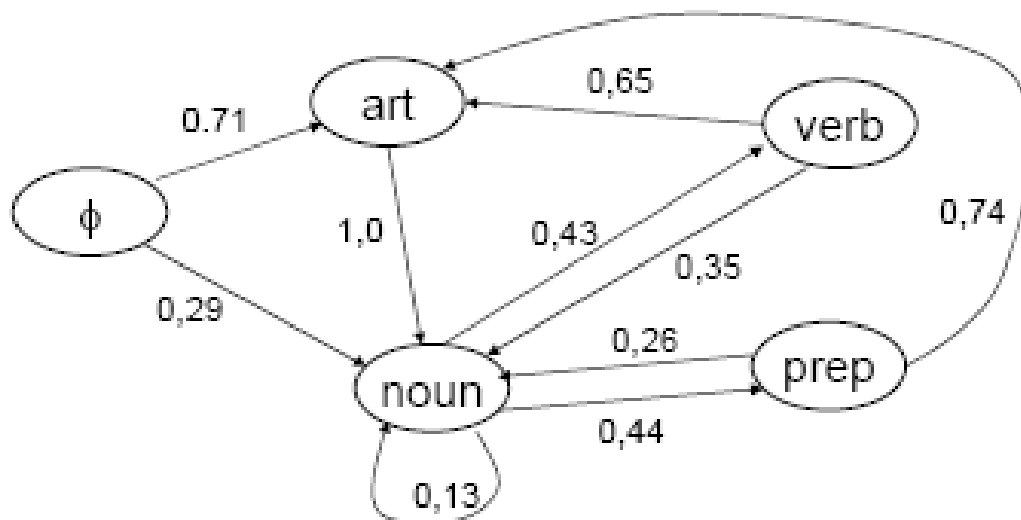
- Οι πιθανότητες $P(w_i | c_i)$ υπολογίζονται μετρώντας απλά το πλήθος των λέξεων που εμφανίζονται υπό συγκεκριμένη κατηγορία
- Παράδειγμα:
 - *John went to the river. He found his rod. A large fish swam past him. He caught the fish and ate it for his tea.*
 - $P(\text{fish} | \text{noun}) = 2 / 5$

Ένας απλός tagger

- Για την ανάλυση μιας πρότασης
 - Δημιούργησε όλες τις δυνατές ακολουθίες κατηγοριών
 - Για την κάθε ακολουθία υπολόγισε την πιθανότητα εμφάνισης με βάση το corpus
 - Διάλεξε την καλύτερη ακολουθία
 - *The fly flies* {
 - art noun noun
 - art noun verb
 - art verb noun
 - art verb verb
 - Αν υπάρχουν T λέξεις και κάθε λέξη έχει N κατηγορίες, τότε απαιτούνται N^T υπολογισμοί

Αλυσίδες Markov (Markov Chains)

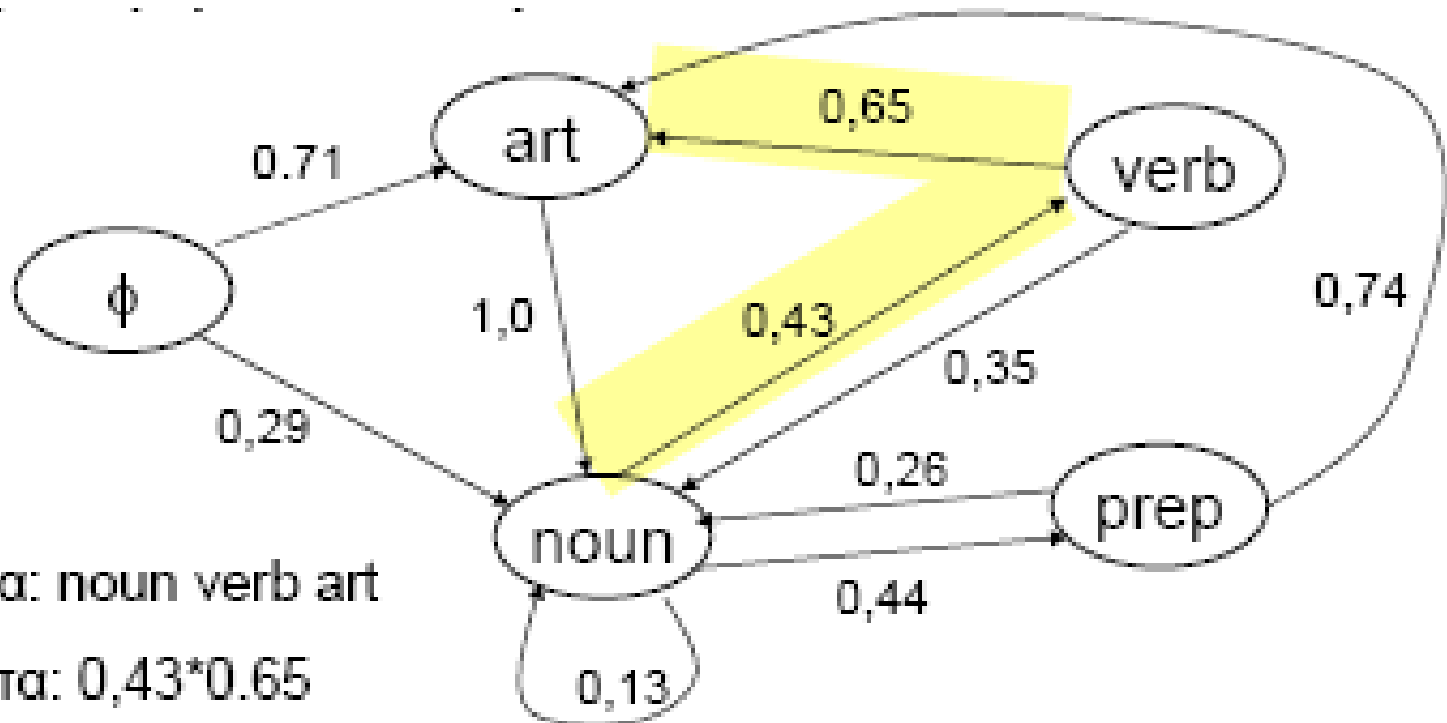
- Οι πιθανότητες $P(c_i | c_{i-1})$ μπορούν να αναπαρασταθούν ως ένα ειδικό δίκτυο μεταβάσεων που καλείται αλυσίδα Markov
- Το άθροισμα των πιθανοτήτων που εξέρχονται από ένα κόμβο πρέπει να είναι 1.



Υπολογισμός πιθανοτήτων

Η πιθανότητα μιας ακολουθίας κατηγοριών είναι το γινόμενο των μονοπατιών

Αυτή η προσέγγιση αγνοεί τις πιθανότητες των συγκεκριμένων λέξεων



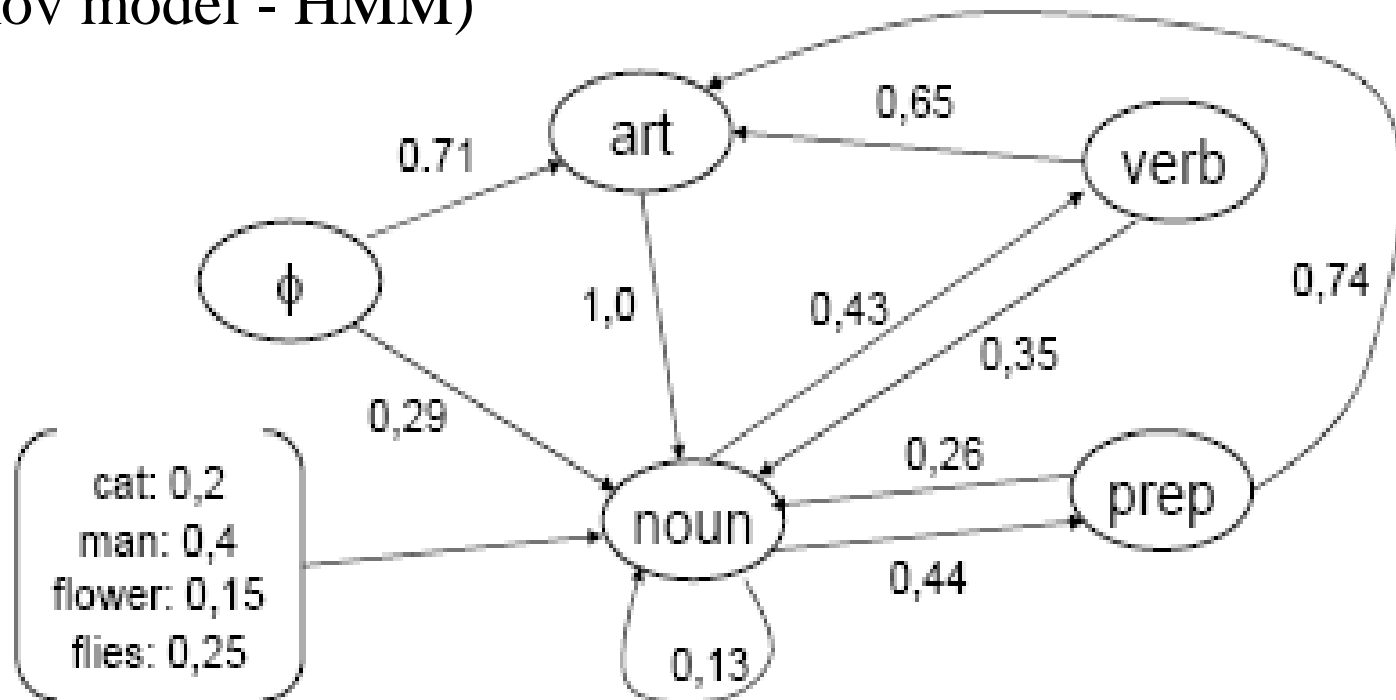
Ακολουθία: noun verb art

Πιθανότητα: $0,43 \cdot 0,65$

Πρόσθεση λεξιλογικών πιθανοτήτων

Τις πιθανότητες $P(w_i | c_i)$ μπορούμε να τις προσθέσουμε ως πίνακα σε κάθε κόμβο

Το μοντέλο αυτό ονομάζεται κρυμμένο μοντέλο Markov (Hidden Markov model - HMM)



Αναγνώριση φράσεων (Text Chunking)

- Χωρισμός μιας πρότασης σε μη επικαλυπτόμενα τμήματα βάσει μιας απλής συντακτικής ανάλυσης
- Ανίχνευση βασικών φράσεων
 - ονοματικών
 - ρηματικών
 - προθετικών
 - επιρρηματικών
- Προπομπός full-parsing και περαιτέρω ανάλυσης

Chunking: Παράδειγμα

- Εύρεση βασικών ονοματικών φράσεων
- Προσκόληση προσδιορισμών σε βασικές ονοματικές φράσεις, εύρεση ρηματικών και άλλων τύπων φράσεων
 - [N Some bankers N] [V are reporting V] [N more inquiries than usual N] [ADV since Friday ADV] .

Το Chunking ως διαδικασία Tagging

- Στόχος είναι να αποδοθεί σε κάθε λέξη ένα tag που υποδεικνύει αν είναι στην αρχή ή εντός μίας φράσης
- Σύμβολα ορίων:
 - B: αρχή
 - I: εντός
 - O: εκτός
- Πλήρες tagset:
 - NP-B: αρχή ονοματικής φράσης
 - NP-I: εντός ονοματικής φράσης
 - VP-B: αρχή ρηματικής φράσης
 - VP-I: εντός ρηματικής φράσης
 - ...

Chunking ως Tagging: Παράδειγμα

- [N Some bankers N] [V are reporting V] [N more inquiries than usual N] [ADV since Friday ADV] .
- Some/NP-B bankers/NP-I are/VP-B reporting/VP-I more/NP-B inquiries/NP-I than/NP-I usual/NP-I since/NP-B Friday/NP-I ./O

Προσάρτηση Προθετικών Φράσεων (PP attachment)

- “Βλέπω τον άνθρωπο με το τηλεσκόπιο”
- Η φράση «με το τηλεσκόπιο» προσαρτάται στο ρήμα ή στο αντικείμενο της πρότασης;

Λέξη	Συχν. εμφάνισης λέξης	Συχν. συνεμφάνισης λέξης + «με»	P(“με” λέξη)
βλέπω	5156	607	$607/5156=0.118$
άνθρωπο	1442	155	$155/1442=0.107$

Στοχαστικές Γραμματικές

- Η θεωρία πιθανοτήτων μπορεί να βελτιώσει τις γραμματικές που έχουμε κατασκευάσει χειρονακτικά
- Μία απλή προσέγγιση είναι να καθορίσουμε μία πιθανότητα για τον κάθε κανόνα γραμματικής με βάση το πόσες φορές χρησιμοποιείται σε ένα σχολιασμένο corpus
- Στοχαστικές γραμματικές ελεύθερης σύνταξης (Probabilistic Context-free Grammars - PCFGs)

Τυπικός Ορισμός μιας PCFG

- Μία PCFG αποτελείται από:
 - Ένα σύνολο τερματικών συμβόλων V_T
 - Ένα σύνολο μη-τερματικών συμβόλων V_N έτσι ώστε $V_T \cap V_N = \emptyset$
 - Μία μεταβλητή έναρξης S
 - Ένα σύνολο κανόνων της μορφής $A \rightarrow B$, όπου $A \in V_N$ και B μία ακολουθία τερματικών και μη-τερματικών συμβόλων
 - Ένα αντίστοιχο σύνολο πιθανοτήτων για τους κανόνες έτσι ώστε για κάθε i $\sum_k P(A_i \rightarrow B_k) = 1$

Στοχαστικές Γραμματικές: Παράδειγμα

□ Corpus:

John talks. He laughs. He likes Mary. He puts the book on the shelf.

□ Γραμματική:

VP \rightarrow V (1)

VP \rightarrow V NP (2)

VP \rightarrow V NP PP (3)

□ Πιθανότητες: (1) 0.5

(2) 0.25

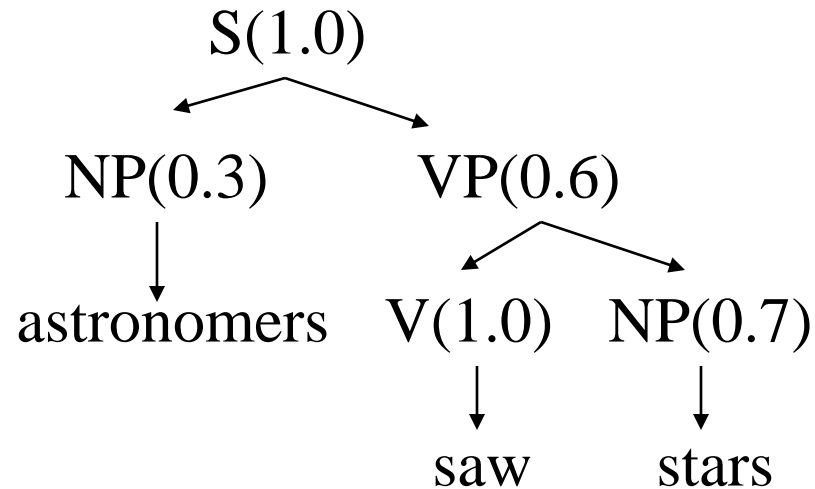
(3) 0.25

Στοχαστικές Γραμματικές: Ανάλυση

- Μπορούμε να θεωρήσουμε ότι η πιθανότητα μετά την εφαρμογή ενός κανόνα είναι το γινόμενο της πιθανότητας του κανόνα επί τις πιθανότητες των συστατικών του δεξιού μέρους του κανόνα

- | | |
|------------------------------------|-------|
| $S \rightarrow NP VP$ | (1.0) |
| $PP \rightarrow P NP$ | (1.0) |
| $VP \rightarrow V$ | (0.4) |
| $VP \rightarrow V NP$ | (0.6) |
| $NP \rightarrow N$ | (1.0) |
| $V \rightarrow \text{saw}$ | (1.0) |
| $N \rightarrow \text{astronomers}$ | (0.3) |
| $N \rightarrow \text{stars}$ | (0.7) |

“Astronomers saw stars”



$$P(\text{“astronomers saw stars”}) = 1.0 * 0.3 * 0.6 * 1.0 * 0.7 = 0.126$$

Deduction (Συναγωγή)-Induction (Επαγωγή): Πλεονεκτήματα

- Deduction:
 - Η γλωσσολογική γνώση και διαίσθηση που έχουμε μπορεί να χρησιμοποιηθεί
 - Ακρίβεια
- Induction:
 - Γρηγορότερη η υλοποίηση του μοντέλου
 - Καλή κάλυψη
 - Robustness
 - Μικρές απαιτήσεις σε γνώση (knowledge poor)
 - Εφαρμόσιμη σε πραγματικά δεδομένα
 - Το μοντέλο είναι εύκολα επεκτάσιμο

Deduction (Συναγωγή)-Induction (Επαγωγή): Μειονεκτήματα

□ Deduction:

- Μεγάλο κόστος και χρονική διάρκεια κατασκευής του μοντέλου
- Δύσκολα εφαρμόσιμη σε πραγματικά δεδομένα με καλή κάλυψη
- Το μοντέλο δεν μπορεί να επεκταθεί εύκολα

□ Induction:

- Δεδομένα πολύ χαμηλής συχνότητας (sparse data)
- Δυσκολία υπολογισμού στατιστικής συσχέτισης γλωσσολογικών φαινομένων