



# ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου

[kerman@ionio.gr](mailto:kerman@ionio.gr)

SEQUENCE  
LABELING

TAGGING

# 1. Part of Speech Tagging (PoS Tagging)

## Αναγνώριση Μερών του Λόγου

# Τα Μέρη του Λόγου (Parts of Speech)

- ❖ Από την προ Χριστού εποχή είχαν γίνει προσπάθειες να κατηγοριοποιηθούν οι λέξεις σε γραμματικές κατηγορίες
- ❖ Τα 8 Μέρη του Λόγου (ΜτΛ) του Διονύσιου του Θραξ (1<sup>ος</sup> αιώνας πΧ)
  - ❖ Ουσιαστικό, Ρήμα, Αντωνυμία, Πρόθεση, Επίρρημα, Σύνδεσμος, Μετοχή, Αρθρο
- ❖ Κλειστά ΜΤΛ
  - ❖ Συνήθως είναι λειτουργικές λέξεις (function words): Μικρές λέξεις, εμφανίζονται πολύ συχνά στην γλώσσα, χωρίς σημασιολογικό περιεχόμενο, και έχουν λειτουργικό ρόλο μέσα στην πρόταση
  - ❖ Άρθρα, Αντωνυμίες, Προθέσεις, Σύνδεσμοι
- ❖ Ανοιχτά ΜΤΛ
  - ❖ Λέξεις με σημασιολογικό περιεχόμενο: Ουσιαστικά, Επίθετα, Ρήματα, Επιρρήματα
  - ❖ Η λίστα τους συνεχώς εμπλουτίζεται με καινούριες λέξεις (σκανάρω, viral)

## Open class ("content") words

### Nouns

#### Proper

*Janet*  
*Italy*

#### Common

*cat, cats*  
*mango*

### Verbs

#### Main

*eat*  
*went*

Adjectives *old green tasty*

Adverbs *slowly yesterday*

### Numbers

*122,312*  
*one*

Interjections *Ow hello*

## Closed class ("function")

Determiners *the some*

Conjunctions *and or*

Pronouns *they its*

### Auxiliary

*can*  
*had*

Prepositions *to with*

Particles *off up*

# Τα Μέρη του Λόγου (Parts of Speech)

Δημιουργεί πρόβλημα κυρίως στα Αγγλικά (~60% των λέξεων σε ένα αγγλικό κείμενο είναι αμφίσημες).

Αμφισημία: Η ίδια λέξη μπορεί σε μια γλώσσα να παίρνει περισσότερα από ένα ΜΤΛ (POS tags)

*The **back** door* (επίθετο)

*On my **back*** (ουσιαστικό)

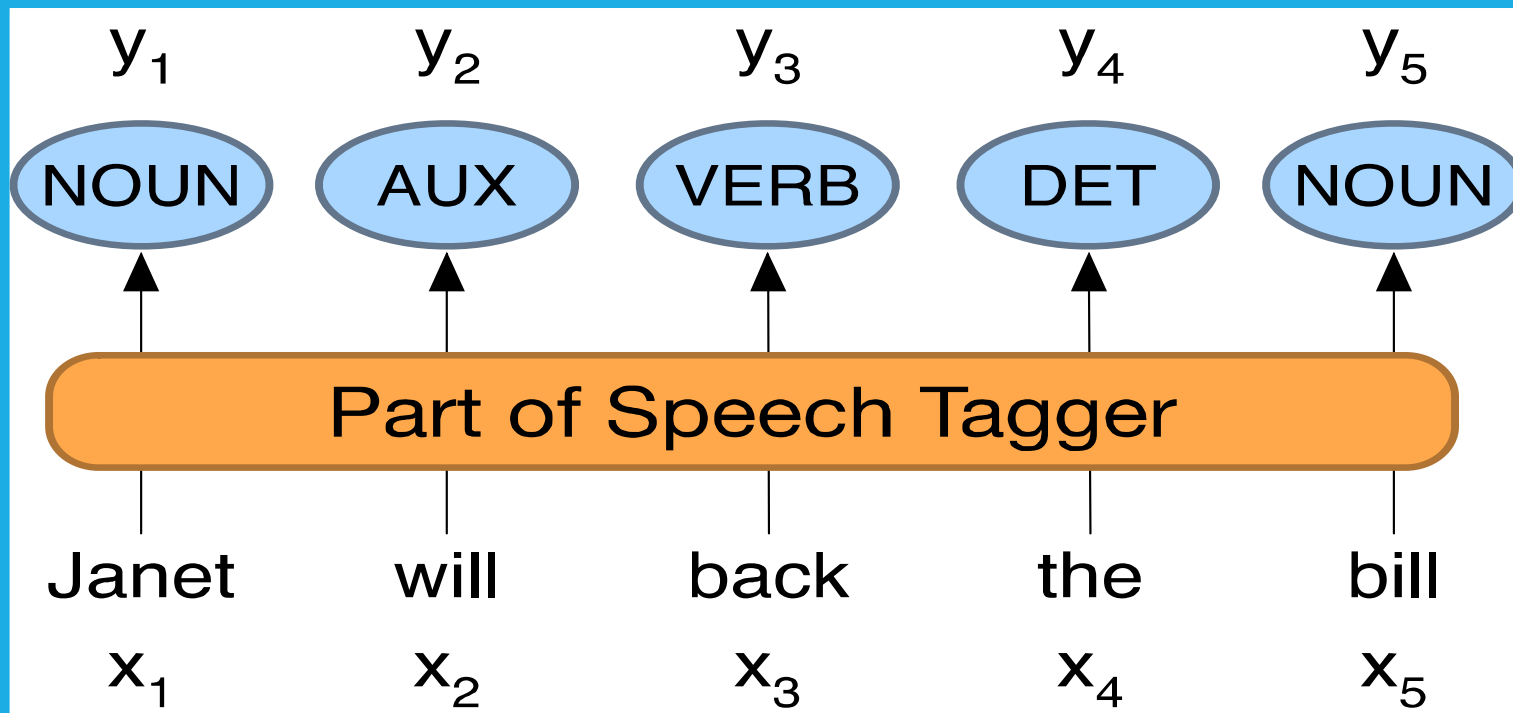
*Win the voters **back*** (μόριο)

*Sit in the **back**.* (επίρρημα)

*Promised to **back** the bill* (ρήμα)

# Part of Speech Tagging

Αντιστοίχιση των λέξεων  $x_1, \dots, x_n$  με τις ετικέτες ΜΤΛ  $y_1, \dots, y_n$



# Παράδειγμα Επισημείωσης MTL

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** ./.

# Χρησιμότητα Επισημείωσης ΜΤΛ

Η γνώση των ετικετών των ΜΤΛ βοηθάει περαιτέρω εργασίες γλωσσικής τεχνολογίας

- ❖ Την συντακτική ανάλυση, όπου πχ ένας συντακτικός κανόνας μπορεί να συνδέει ένα άρθρο και ένα ουσιαστικό σε μια ονοματική φράση
- ❖ Την μετατροπή κειμένου σε ομιλία (TTS) όπου το ΜΤΛ μπορεί να καθορίσει αν θα προφέρω το lead “λιντ” (οδηγώ - ΡΗΜΑ) ή “λεντ” (σίδερο - ΟΥΣΙΑΣΤΙΚΟ)
- ❖ Την αναγνώριση συναισθήματος, όπου πχ πρέπει να αναγνωρίσω τα επίθετα
- ❖ Την Μηχανική Μετάφραση, όπου πχ ένα επίθετο που ακολουθείται από ένα ουσιαστικό στα Ελληνικά θα πρέπει να τα αναδιατάξω στα Ισπανικά ώστε το επίθετο να ακολουθεί το ουσιαστικό
- ❖ Τον υπολογισμό σημασιολογικής ομοιότητας/διαφοράς ανάμεσα σε κείμενα
- ❖ ...

# Η Απόδοση στην Αναγνώριση ΜΤΛ

**Accuracy (Ακρίβεια):** Σε ένα κείμενο δοκιμής πόσες ετικέτες που προβλέφθηκαν είναι σωστές; (μετريέται στις αμφίσημες λέξεις)

## **Unigram model (Baseline)**

Εκπαίδευση: Από ένα μεγάλο σώμα κειμένων, επισημειωμένο ως προς το ΜΤΛ των λέξεών του από ειδικούς, υπολογίζεται το ΜΤΛ με το οποίο εμφανίζεται πιο συχνά.

Δοκιμή: Σε κάθε καινούρια εμφάνιση της λέξης αποδίδεται αυτό το ΜΤΛ.

90% ακρίβεια αν στην κάθε λέξη αποδίδεται πάντα το πιο συχνό tag.

Σύγχρονα state-of-the-art εργαλεία: Ακρίβεια > 99%.

# Πηγές γνώσης για την Αναγνώριση ΜΤΛ

Anna

will

pay the

cook.

ΟΥΣΙΑΣΤΙΚΟ/ΒΟΗΘΗΤΙΚΟ ΡΗΜΑ?

ΟΥΣΙΑΣΤΙΚΟ/ΡΗΜΑ?

Η γνώση που μπορώ να χρησιμοποιήσω για την αναγνώριση του ΜΤΛ είναι

- ❖ Οι εκ των προτέρων πιθανότητες (prior probabilities) μια λέξη να ανήκει σε ένα συγκεκριμένο ΜΤΛ σε μια γλώσσα
  - ❖ Το 'will' είναι συνήθως βοηθητικό ρήμα στα Αγγλικά (unigram)
- ❖ Τα συμφραζόμενα μέσα στα οποία εμφανίζεται η λέξη
  - ❖ Μετά το 'the' συχνά ακολουθεί ουσιαστικό – σπάνια ακολουθεί ρήμα.
- ❖ Η δομή της λέξης
  - ❖ Η "Anna" ξεκινάει με κεφαλαίο – αυτό είναι μεγάλη ένδειξη κύριου ονόματος.
  - ❖ Επιθέματα: ...-ly ΕΠΙΡΡΗΜΑ
  - ❖ Προθέματα: un-... ΕΠΙΘΕΤΟ

# POS Tagsets για τα Αγγλικά (Wall Street Journal)

|     |                   |                     |      |                     |                    |
|-----|-------------------|---------------------|------|---------------------|--------------------|
| 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>+, %, &amp;</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>          |
| VCN | 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>   |      |                  |                   |

# Σώματα Κειμένων

Όλες οι μέθοδοι αναγνώρισης MTL στηρίζονται στην ανάλυση μεγάλων χειρωνακτικά επισημειωμένων σωμάτων κειμένων.

Στις στατιστικές μεθόδους η ανάλυση είναι εξαγωγή μετρήσεων (πχ *fly*: είναι ουσιαστικό στο 95% των περιπτώσεων που προηγείται άρθρο - *the fly*). Αυτές οι μετρήσεις χρησιμοποιούνται για την απόδοση MTL σε λέξεις καινούριων προτάσεων.

Όλες οι στατιστικές μέθοδοι στηρίζονται στο θεώρημα του Bayes

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

$$P(c_1..c_n \mid w_1..w_n) = P(w_1..w_n \mid c_1..c_n) * P(c_1..c_n) / P(w_1..w_n)$$

Σε μία ακολουθία λέξεων  $w_1..w_n$  θέλουμε να αποδώσουμε την πιο πιθανή ακολουθία MTL (κατηγοριών)  $c_1..c_n$  που μεγιστοποιεί την παραπάνω πιθανότητα  
πχ  $P(\text{art noun verb} \mid \text{the fly flies})$

# Στατιστική Αναγνώριση ΜΤΛ

$$P(c_1..c_n \mid w_1..w_n) = P(w_1..w_n \mid c_1..c_n) * P(c_1..c_n) / P(w_1..w_n)$$

Πιθανότητα να εμφανιστεί η συγκεκριμένη ακολουθία λέξεων στην γλώσσα μου δεδομένου ότι ανήκουν σε αυτή την ακολουθία μερών του λόγου

Πιθανότητα να εμφανιστεί η συγκεκριμένη ακολουθία ΜΤΛ στην γλώσσα μου

Πιθανότητα να εμφανιστεί η συγκεκριμένη ακολουθία λέξεων στην γλώσσα μου (ΜΟΝΤΕΛΟ ΓΛΩΣΣΑΣ)

$w_1$   $w_2$   $w_3$   
the fly flies

|   | c1         | c2   | c3   |
|---|------------|------|------|
| { | art        | noun | noun |
|   | <u>art</u> | noun | verb |
|   | art        | verb | noun |
|   | art        | verb | verb |

**Ποια από τις 4 πιθανότητες είναι η μεγαλύτερη;**



$$\begin{aligned} P(\text{art noun noun} \mid \text{the fly flies}) &= P(\text{the fly flies} \mid \text{art noun noun}) * P(\text{art noun noun}) / P(\text{the fly flies}) \\ P(\text{art noun verb} \mid \text{the fly flies}) &= P(\text{the fly flies} \mid \text{art noun verb}) * P(\text{art noun verb}) / P(\text{the fly flies}) \\ P(\text{art verb noun} \mid \text{the fly flies}) &= P(\text{the fly flies} \mid \text{art verb noun}) * P(\text{art verb noun}) / P(\text{the fly flies}) \\ P(\text{art verb verb} \mid \text{the fly flies}) &= P(\text{the fly flies} \mid \text{art verb verb}) * P(\text{art verb verb}) / P(\text{the fly flies}) \end{aligned}$$

# Απλοποίηση του στατιστικού μοντέλου

$$P(c_1..c_n \mid w_1..w_n) = P(w_1..w_n \mid c_1..c_n) \quad * \quad P(c_1..c_n) \quad / \quad P(w_1..w_n)$$

 Παραδοχή ανεξαρτησίας των δεδομένων (Conditional Independence Assumption): Σε μια ακολουθία λέξεων κάθε λέξη είναι ανεξάρτητη των υπολοίπων λέξεων.

$$P(w_1..w_n \mid c_1..c_n) = \prod_{i=1..n} P(w_i \mid c_i)$$

 Μοντέλο N-γράμμου (N-gram model): Κάθε MTL εξαρτάται από το MTL των n-1 προηγούμενων λέξεων στην ακολουθία.

$$P(c_1..c_n) =$$

$$\prod_{i=1..n} P(c_i \mid c_1, \dots, c_{n-1})$$

Μοντέλο bigram

$$P(c_1..c_n) = \prod_{i=1..n} P(c_i \mid c_{i-1})$$

 Αν δούμε τον υπολογισμό των προηγούμενων 4 πιθανοτήτων, ο παρονομαστής αυτός δεν αλλάζει – άρα δεν επηρεάζει την κατάταξη των 4 αποτελεσμάτων. Μπορώ να τον αγνοήσω.

**Απλοποιημένο Μοντέλο:**  $P(c_1..c_n \mid w_1..w_n) = \prod_{i=1..n} P(w_i \mid c_i) * \prod_{i=1..n} P(c_i \mid c_{i-1})$

# Απλοποίηση του στατιστικού μοντέλου

$$\text{Απλοποιημένο Μοντέλο: } P(c_1..c_n \mid w_1..w_n) = \prod_{i=1..n} P(w_i \mid c_i) * \prod_{i=1..n} P(c_i \mid c_{i-1})$$

Όλες αυτές οι πιθανότητες υπολογίζονται από το επισημειωμένο σώμα κειμένων ως εξής:



Lexical probabilities  
Emission probabilities

bigram probabilities  
Transition probabilities

- ❖ Οι λεξικολογικές πιθανότητες υπολογίζονται διαιρώντας την συχνότητα εμφάνισης μιας λέξης που ανήκει σε ένα συγκεκριμένο ΜΤΛ με την συνολική εμφάνιση συχνότητα εμφάνισης του ΜΤΛ στο κείμενο  
*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**.*

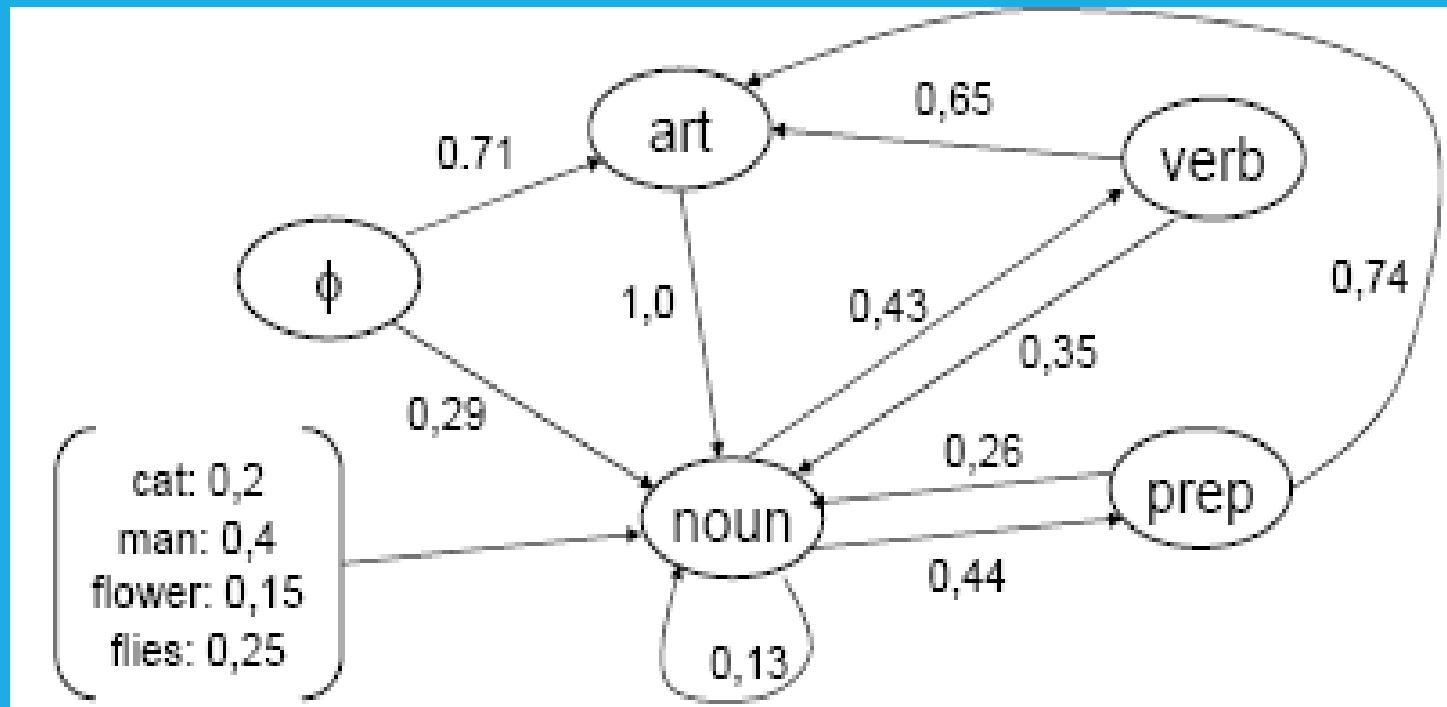
$$P(\text{fish} \mid \text{NOUN}) = 2 / 5$$

- ❖ Οι πιθανότητες διγράμμου υπολογίζονται με **Εκτιμήσεις Μέγιστης Πιθανοφάνειας** (Maximum Likelihood Estimates)

$$P(c_i = \text{VERB} \mid c_{i-1} = \text{NOUN}) = \text{count}(c_i = \text{VERB} \ \& \ c_{i-1} = \text{NOUN}) / \text{count}(c_{i-1} = \text{NOUN})$$

δηλ ο αριθμός ουσιαστικών που ακολουθούνται από ρήματα προς το συνολικό αριθμό των ουσιαστικών  
Ορίζουμε μια επιπλέον κατηγορία 0, για την εκκίνηση μιας πρότασης:

# Η Αναγνώριση ΜΤΛ και τα Κρυμμένα Μοντέλα Markov (Hidden Markov Models)



## 2. IOB Tagging - Chunking

### Αναγνώριση φράσεων

# Αναγνώριση φράσεων

- ❖ 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: ένα πρόβλημα δυο φάσεων

- ❖ Φάση Πρώτη (Bracketing): Αναγνώριση ορίων των φράσεων
- ❖ Φάση Δεύτερη (Tagging): Αναγνώριση του είδους της κάθε φράσης

[N Some bankers N] [V are reporting V] [N more inquiries than usual N] [ADV since Friday ADV] .

# Chunking: Μετατροπή σε πρόβλημα Tagging (IOB Tagging)

Στόχος είναι να αποδοθεί σε κάθε λέξη ένα tag που υποδεικνύει αν είναι στην αρχή ή εντός μίας φράσης και τι είδους φράσης

Σύμβολα ορίων:

B: αρχή φράσης

I: εντός φράσης

O: εκτός οποιασδήποτε φράσης

Πλήρες tagset:

NP-B: αρχή ονοματικής φράσης

NP-I: εντός ονοματικής φράσης

VP-B: αρχή ρηματικής φράσης

VP-I: εντός ρηματικής φράσης

...

[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

### 3. Named Entity Recognition (NER)

### Αναγνώριση Ονομάτων

# Named Entities – Κύρια Ονόματα

Οτιδήποτε μπορεί να εκφραστεί σαν Κύριο Όνομα. Μπορεί να είναι μια λέξη ή φράσεις περισσότερων λέξεων (Multi-word expressions)

Οι πιο συχνές κατηγορίες κυρίων ονομάτων είναι

- ❖ Ονόματα ανθρώπων (PER): «Ιωάννης Θεοτόκης»
- ❖ Τοπωνύμια (LOC): «Άστρο Κυνουρίας»
- ❖ Ονόματα Οργανισμών (ORG): «Ιόνιο Πανεπιστήμιο»
- ❖ Πολιτικές Οντότητες (GPE): «Νέα Δημοκρατία»
- ❖ Άλλες κατηγορίες είναι
  - ❖ Ημερομηνίες, Τιμές, Ποσά, Προϊόντα, Επιστημονικοί Όροι (πχ Βιοχημικοί, Ιατρικοί, Νομικοί, Οικονομικοί κλπ)

# NER: ένα πρόβλημα δυο φάσεων

- ❖ Φάση Πρώτη (Bracketing): Αναγνώριση ορίων των Ονομάτων
- ❖ Φάση Δεύτερη (Tagging): Αναγνώριση της κατηγορίας του κάθε ονόματος

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **\$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

# Χρησιμότητα του NER

- ❖ Ανάλυση Συναισθήματος: Θέλω το συναίσθημα για ένα προϊόν, πρόσωπο, εταιρία
- ❖ Απάντηση Ερωτημάτων (Question Answering): Θέλω απάντηση σχετικά με μια οντότητα
- ❖ Εξαγωγή Πληροφορίας (Information Extraction): Θέλω να βγάλω από ένα κείμενο πληροφορία σχετικά με μια οντότητα

# Αμφισημία στο NER

- ❖ Το ίδιο Όνομα μπορεί να εμφανίζεται στη γλώσσα με περισσότερες από μια κατηγορίες NER.

[PER Washington] was born into slavery on the farm of James Burroughs.

[ORG Washington] went up 2 games to 1 in the four-game series.

Blair arrived in [LOC Washington] for what may well be his last state visit.

In June, [GPE Washington] passed a primary seatbelt law.

# Και πάλι IOB: Μετατροπή του NER σε πρόβλημα Tagging

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago ] route.

| Words      | BIO Label |
|------------|-----------|
| Jane       | B-PER     |
| Villanueva | I-PER     |
| of         | O         |
| United     | B-ORG     |
| Airlines   | I-ORG     |
| Holding    | I-ORG     |
| discussed  | O         |
| the        | O         |
| Chicago    | B-LOC     |
| route      | O         |
| .          | O         |

# Επέκταση του IOB: το σχήμα IOBES

[PER Jane Villanueva]  
of [ORG United] , a unit  
of [ORG United Airlines  
Holding] , said the fare  
applies to the [LOC  
Chicago ] route.

E: τέλος NE

S: NE μίας λέξης

| Words      | IO Label | BIO Label | BIOES Label |
|------------|----------|-----------|-------------|
| Jane       | I-PER    | B-PER     | B-PER       |
| Villanueva | I-PER    | I-PER     | E-PER       |
| of         | O        | O         | O           |
| United     | I-ORG    | B-ORG     | B-ORG       |
| Airlines   | I-ORG    | I-ORG     | I-ORG       |
| Holding    | I-ORG    | I-ORG     | E-ORG       |
| discussed  | O        | O         | O           |
| the        | O        | O         | O           |
| Chicago    | I-LOC    | B-LOC     | S-LOC       |
| route      | O        | O         | O           |
| .          | O        | O         | O           |

## 4. Άρση Αμφισημίας Λέξεων – Word Sense Disambiguation (WSD)

# Η Αμφισημία στις έννοιες των λέξεων

Πολλές λέξεις έχουν αρκετές διαφορετικές έννοιες

*Ποντίκι*: τρωκτικό, ηλ.συσκευή, μέρος κρέατος

*Γέφυρα*: κατασκευή, προσθετικό οδοντικής

*Βιβλιοθήκη*: κτίριο, έπιπλο

Συχνά η έννοια μιας λέξης γίνεται σαφής από τα συμφραζόμενα μιας πρότασης

*Έπιασα στη φάκα ένα ποντίκι.* (τρωκτικό)

*Αγόρασα ένα ασύρματο ποντίκι.* (συσκευή)

*Αγόρασα ένα κιλό ποντίκι.* (κρέας)

*Αγόρασα ένα λευκό ποντίκι.* (τρωκτικό, συσκευή)

# Χρησιμότητα του WSD

Στην μηχανική μετάφραση

We played at the river **bank** → Παίξαμε στην **όχθη** του ποταμού

He opened an account at the local **bank** → Άνοιξε λογαριασμό στην **τράπεζα** της περιοχής

Στην εξαγωγή πληροφορίας

Ένα σύστημα ΕΠ πρέπει να επιστρέφει κείμενα σχετικά με τράπεζες σε μια ερώτηση που περιλαμβάνει τους όρους 'financial bank'

# Στατιστική Άρση της Αμφισημίας: Μοντέλο Μονογράμμου (Unigram Model)

Η πιο απλή στατιστική προσέγγιση είναι να μετρήσουμε πόσες φορές χρησιμοποιείται μία λέξη με την κάθε δυνατή έννοια που μπορεί να πάρει στη γλώσσα μέσα σε ένα corpus

γέφυρα1 (πρ. οδοντικής): 221 φορές

γέφυρα2 (κατασκευή): 4356 φορές

Αυτές οι μετρήσεις αναφέρονται ως unigrams

Απαιτούν την ύπαρξη κατάλληλα σχολιασμένου corpus

**Διαλέγω πάντα την πιο συχνή έννοια με την οποία εμφανίζεται η λέξη στη γλώσσα μου, και αυτή την έννοια της αποδίδω.**

# Στατιστική Άρση της Αμφισημίας: Μοντέλο N-γράμμου (N-gram Model)

Πρέπει να λάβω υπόψη μου τα συμφραζόμενα

Αν  $s_i$  είναι η έννοια (sense) της λέξης  $i$

Bigrams:  $P(s_n \mid s_{n-1})$

Trigrams:  $P(s_n \mid s_{n-1}, s_{n-2})$

Θεωρούμε ένα παράθυρο λέξεων στο οποίο η λέξη που μας ενδιαφέρει βρίσκεται στη μέση

$w_1..w_5..w_9$  όπου μας ενδιαφέρει η έννοια της λέξης  $w_5$

Θέλουμε να βρούμε την έννοια  $s$  της λέξης  $w_5$  ( $w_5/s$ ) που μεγιστοποιεί τη πιθανότητα  $P(w_5/s \mid w_1..w_9)$

# Στατιστική Άρση της Αμφισημίας: Απλοποίηση του Μοντέλου (Naïve Bayes)

- Bayes rule:  $P(A|B)=P(A)*P(B|A)/P(B)$
- Αν  $w$  είναι η κεντρική λέξη (λέξη εστίασης) του παραθύρου  $w_1..w_n$
- $P(w/s \mid w_1..w_n) = P(w/s)*P(w_1..w_n \mid w/s) / P(w_1..w_n)$

πιθανότητα έννοιας λέξης  
δεδομένων των  
συμφραζομένων (posterior)

πιθανότητα έννοιας  
λέξης ανεξάρτητη των  
συμφραζομένων (prior)

παράγοντας  
ενσωμάτωσης  
πληροφορίας  
συμφραζομένων

- Το  $P(w_1..w_n)$  είναι σταθερό
- Παραδοχή ανεξαρτησίας (η παρουσία μιας λέξης στα συμφραζόμενα είναι ανεξάρτητη των άλλων λέξεων):

$$P(w_1..w_n \mid w/s) \rightarrow \prod_{i=1..n} P(w_i \mid w/s)$$

- Τελικό μοντέλο:  $P(w/s \mid w_1..w_n)=P(w/s)* \prod_{i=1..n} P(w_i \mid w/s)$

# Στατιστικό Μοντέλο: POS Tagging vs WSD

- ❑ Υπάρχουν πολύ περισσότερες σημασιολογικές έννοιες από κατηγορίες Μερών του Λόγου
- ❑ Στην πραγματικότητα δεν μπορούμε να ορίσουμε πλήρως όλες τις επιτρεπτές σημασιολογικές ιδιότητες μιας έννοιας
  - ❑ Έξυπνος άνθρωπος
  - ❑ Έξυπνη συσκευή
  - ❑ Έξυπνες κάρτες
  - ❑ Έξυπνα πλυντήρια
- ❑ Ο αριθμός των φορών που εμφανίζεται η κάθε έννοια μπορεί να είναι πάρα πολύ μικρός
- ❑ Το «παράθυρο» των συμφραζομένων πρέπει να είναι πολύ μεγαλύτερο
  - ❑ Ο Γιάννης πήγε στον οδοντίατρο για να του φτιάξει τη γέφυρα.
  - ❑ Η σημαντική λέξη βρίσκεται 6 λέξεις μακριά
  - ❑ Στο παράθυρο δεν μετράω λέξεις λειτουργικές που δεν έχουν σημασιολογικό περιεχόμενο (κρατάω μόνο ουσιαστικά, ρήματα, επίθετα, επιρρήματα)

# Στατιστική Άρση της Αμφισημίας: Εξαγωγή Μετρήσεων

- $P(w/s \mid w_1..w_n) = P(w/s) * \prod_{i=1..n} P(w_i \mid w/s)$
- Έστω ότι από ένα σχολιασμένο σώμα κειμένων μετράω πιθανότητες εμφάνισης μιας αμφίσημης λέξης (λέξη εστίασης) με μια συγκεκριμένη έννοια  $P(w/s)$ 
  - $P(\text{γέφυρα}|s) = \text{πόσες φορές εμφανίζεται στο σώμα η λέξη } \text{γέφυρα} \text{ με την έννοια } s / \text{πόσες φορές εμφανίζεται στο σώμα η } \text{γέφυρα}$
  - $P(\text{γέφυρα}/\text{κατασκευή οδοποιίας}) = 0.7$   $P(\text{γέφυρα}/\text{πρ. οδοντικής}) = 0.3$
- Μετράω πιθανότητες εμφάνισης μιας λέξης που βρίσκεται μέσα στο παράθυρο συμφραζομένων της λέξης εστίασης δεδομένου ότι η λέξη εστίασης έχει μια συγκεκριμένη έννοια  $s$   $P(w_i \mid w/s)$ 
  - $P(\text{οδοντίατρος} \mid \text{γέφυρα}/\text{πρ. οδοντικής}) = \text{πόσες φορές εμφανίζεται η λέξη } \text{οδοντίατρος} \text{ στο παράθυρο συμφραζομένων της λέξης } \text{γέφυρα} \text{ όταν αυτή έχει έννοια προσθετικού οδοντικής} / \text{πόσες φορές εμφανίζεται η λέξη } \text{οδοντίατρος} \text{ στο παράθυρο συμφραζομένων της λέξης } \text{γέφυρα}$

# Στατιστική Άρση της Αμφισημίας: Παράδειγμα

- $P(w/s \mid w_1..w_n) = P(w/s) * \prod_{i=1..n} P(w_i \mid w/s)$
- $w_1..w_n$  : «**Πήγα να βάλω μία γέφυρα στο μπροστινό μου δόντι.**»
- $P(\text{γέφυρα/κατ. οδοποιίας} \mid w_1..w_n) = P(\text{γέφυρα/κατ. οδοποιίας}) * P(\text{πήγα} \mid \text{γέφυρα/κατ. οδοποιίας}) \times P(\text{βάλω} \mid \text{γέφυρα/κατ. οδοποιίας}) \times P(\text{μπροστινό} \mid \text{γέφυρα/κατ. οδοποιίας}) \times P(\text{δόντι} \mid \text{γέφυρα/κατ. οδοποιίας})$
- $P(\text{γέφυρα/πρ. οδοντικής} \mid w_1..w_n) = P(\text{γέφυρα/ πρ. οδοντικής}) * P(\text{πήγα} \mid \text{γέφυρα/ πρ. οδοντικής}) \times P(\text{βάλω} \mid \text{γέφυρα/ πρ. οδοντικής}) \times P(\text{μπροστινό} \mid \text{γέφυρα/ πρ. οδοντικής}) \times P(\text{δόντι} \mid \text{γέφυρα/ πρ. οδοντικής})$
- Ποια από τις δυο πιθανότητες είναι μεγαλύτερη;
- Η πάνω πορτοκαλί πιθανότητα θα είναι λογικά πολύ πολύ μικρότερη από την κάτω. Οπότε θα είναι αυτή που θα μου θέσει την κάτω πιθανότητα πολύ μεγαλύτερη από την πάνω, και θα επιλύσει καταλυτικά την αμφισημία.

# Άρση της Αμφισημίας: Λίστες Απόφασης (Decision Lists)

- ❑ Πρόκειται για ένα σύνολο κανόνων (ελέγχων) που αφορούν στα συμφραζόμενα της αμφίσημης λέξης μου
- ❑ Ελέγχω έναν-έναν τους κανόνες ξεκινώντας από τον πρώτο
- ❑ Ο πρώτος κανόνας που θα ικανοποιηθεί καθορίζει την έννοια της λέξης
- ❑ Στο τέλος υπάρχει ένας default κανόνας που τρέχει σε περίπτωση που όλοι οι από πάνω κανόνες αποτύχουν
  - ❑ Συνήθως αυτός επιλέγει την πιο συχνή έννοια της λέξης
- ❑ Παράδειγμα: Αποσαφήνιση της έννοιας *γέφυρα*
- ❑ *Λίστα κανόνων*
  1. «κρεμαστή γέφυρα» → κατασκευή
  2. «δόντι» μέσα στο παράθυρο → πρ.οδοντικής
  3. «πυλώνας» μέσα στο παράθυρο → κατασκευή
  4. «οδοντίατρος» μέσα στο παράθυρο → πρ.οδοντικής
  5. Η πιο συχνή έννοια του «γέφυρα» στη γλώσσα

# Λίστες Απόφασης: Μάθηση

- ❑ Οι κανόνες ταξινομούνται κατά φθίνουσα σειρά βάσει ενός σκορ
- ❑ Το σκορ αυτό είναι η πιθανότητα ή λέξη μου να έχει μια συγκεκριμένη έννοια δεδομένου ότι ο κανόνας ικανοποιείται  $P(w/s \mid \text{test})$
- ❑  $P(w/s \mid \text{test}) = \frac{\text{πόσες φορές ικανοποιείται ο κανόνας με την λέξη } w \text{ να έχει την έννοια } s}{\text{πόσες φορές ικανοποιείται ο κανόνας με την λέξη } w}$
- ❑ *Το σκορ αυτό υπολογίζεται αυτόματα από το σώμα εκπαίδευσης, και έτσι οι κανόνες μπορούν να μαθευτούν αυτόματα.*
- ❑ «κρεμαστή γέφυρα» → κατασκευή
  - ❑ Σκορ του παραπάνω κανόνα:
  - ❑ Πόσες φορές εμφανίζεται η λέξη *κρεμαστή* πριν από τη λέξη *γέφυρα* όταν η τελευταία είναι κατασκευή οδοποιίας / Πόσες φορές εμφανίζεται η λέξη *κρεμαστή* πριν από τη λέξη *γέφυρα*

# Άρση Αμφισημίας Έννοιας: Αξιολόγηση

- ❑ Ακρίβεια (precision): Το ποσοστό των λέξεων που μαρκάρονται με τη σωστή έννοια
- ❑ Κάτω όριο: Η ακρίβεια όταν διαλέγουμε πάντα την πιο συχνή έννοια
- ❑ Πάνω όριο: Η ακρίβεια ενός ανθρώπου
- ❑ Η αξιολόγηση απαιτεί ένα σχολιασμένο corpus για εξαγωγή της ακρίβειας
- ❑ Εξαρτάται οι πιθανές έννοιες μιας λέξης πόσο χοντροκομμένες είναι (ευδιάκριτες) ή πόσο εξειδικευμένες είναι (δύσκολα διακρίνονται μεταξύ τους)
  - Ευδιάκριτες, βασικές έννοιες: 95% συμφωνία μεταξύ ανθρώπων
  - Δυσδιάκριτες (εξειδικευμένες) έννοιες: 70% συμφωνία μεταξύ εξειδικευμένων ανθρώπων

**Πολύ πιο πολύπλοκη διαδικασία από το POS tagging**

# Άρση Αμφισημίας Έννοιας: Εξωγενής Αξιολόγηση

- Η αξιολόγηση ενός συστήματος σημασιολογικής αποσαφήνισης μπορεί να γίνει και στα πλαίσια ενός γενικότερου συστήματος
  - Μηχανική μετάφραση
  - Εξαγωγή πληροφορίας
- Μετράται το πόσο βελτιώνεται η απόδοση όλου του συστήματος
- Δεν απαιτείται ύπαρξη σχολιασμένου corpus