

«Τεχνογλωσσία VIII»

Εξαγωγή πληροφοριών από κείμενα

Σεμινάριο 2: Εξαγωγή Πληροφορίας

Ευάγγελος Καρκαλέτσης, Γεώργιος Πετάσης

Εργαστήριο Τεχνολογίας Γνώσεων & Λογισμικού,
Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”
Τηλ.: 210-6503197, Fax: 210-6532175, {vangelis, petasis}@iit.demokritos.gr
Ακαδημαϊκό Έτος: 2013 – 2014

ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

Από κείμενα: Εισαγωγή και Γενική Επισκόπηση

Οι διαφάνειες αυτής της ενότητας βασίζονται
εν μέρει στο κεφάλαιο 22 του βιβλίου:

«Speech and Language Processing»
των D. Jurafsky και J.H. Martin, 2η έκδοση, Pearson, 2009

Το βιβλίο **δεν απαιτείται** για το μάθημα αυτό.

Αξιοποίηση πληροφορίας από κείμενα

- Διάθεση στους χρήστες, μέσω του διαδικτύου και των ψηφιακών βιβλιοθηκών, τεράστιου όγκου πληροφορίας σε ηλεκτρονική μορφή
- Απαιτείται η ανάπτυξη τεχνικών που να επιτρέπουν στους χρήστες τη λήψη και αξιοποίηση εκείνης μόνο της πληροφορίας που σχετίζεται πραγματικά με τα ενδιαφέροντα και τις ανάγκες τους

Προβλήματα - Απαιτήσεις

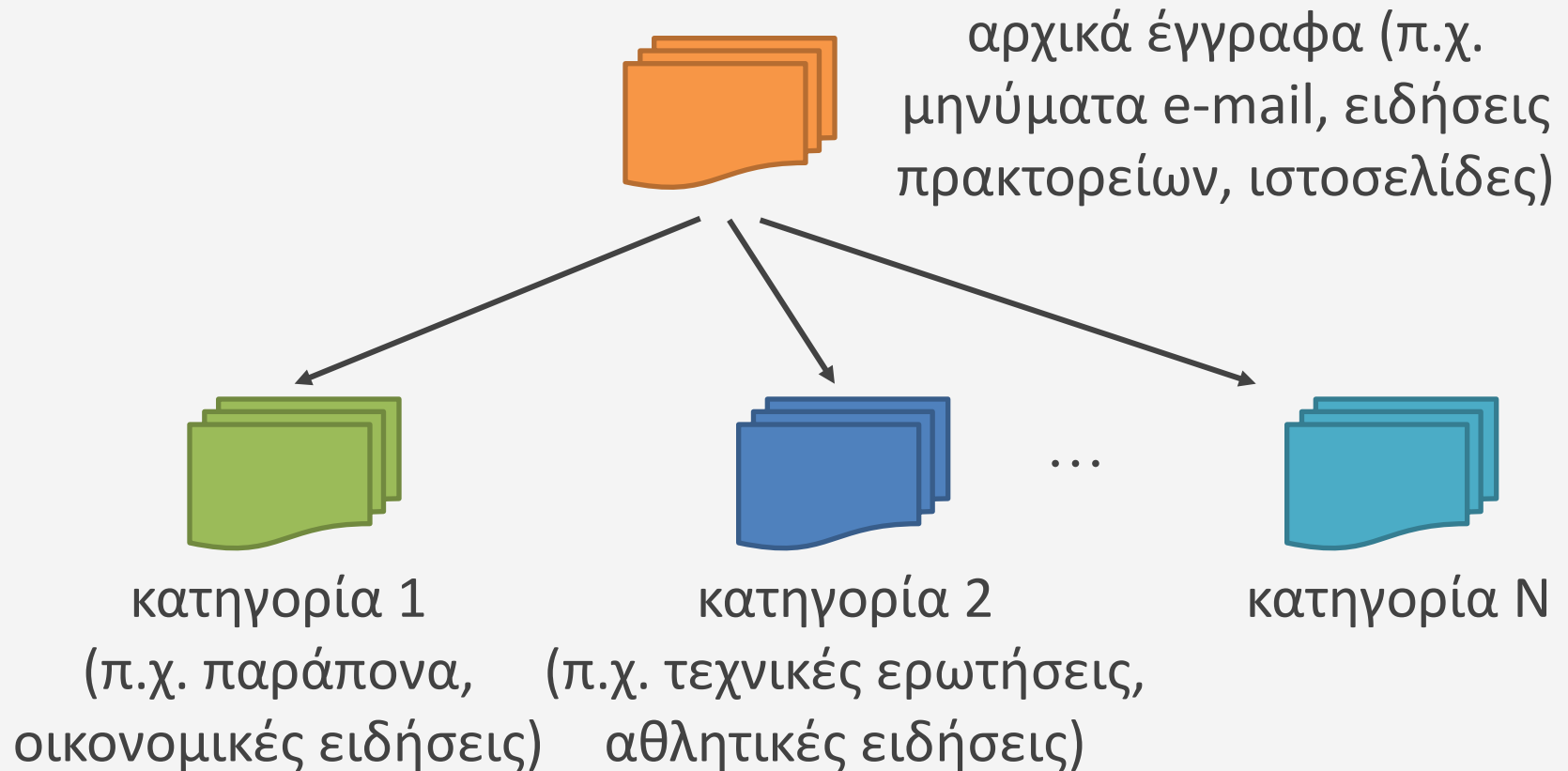
- Περιορισμένη κάλυψη του Παγκόσμιου Ιστού από τις μηχανές αναζήτησης
- Πολύ συχνά, αυτό που πραγματικά ζητάμε δεν κατατάσσεται σε υψηλή θέση από τις μηχανές αναζήτησης
- Διάχυση επιβλαβούς πληροφορίας
- Παροχή ανεπαρκούς πληροφορίας
- Αναζήτηση πληροφορίας σε άλλες γλώσσες
- ...

χρειαζόμαστε πιο έξυπνες προσεγγίσεις!

Ανάγκη για νέες τεχνολογίες

- Ανάκτηση πληροφορίας
- Κατηγοριοποίηση πληροφορίας
- Εξαγωγή πληροφορίας
- Σύντηξη πληροφορίας από πολλαπλά μέσα
- Μοντελοποίηση Χρηστών

Τεχνολογίες: Κατηγοριοποίηση πληροφορίας



Τεχνολογίες: Εξαγωγή πληροφορίας

ΝΕΑ ΕΞΑΓΟΡΑ ΑΠΟ ΤΗ ΣΕΛΟΝΤΑ Ο Όμιλος Σελόντα Riopresca προχώρησε στην εξαγορά έναντι 280.000 € του 80% των μετοχών της Τρίτων ΑΕ Ιχθυοκαλλιέργειας που εδρεύει στο νομό Θεσπρωτίας. Η ετήσια παραγωγή της «Τρίτων» ανέρχεται σε 150 τόνους λαυράκι-τσιπούρα.



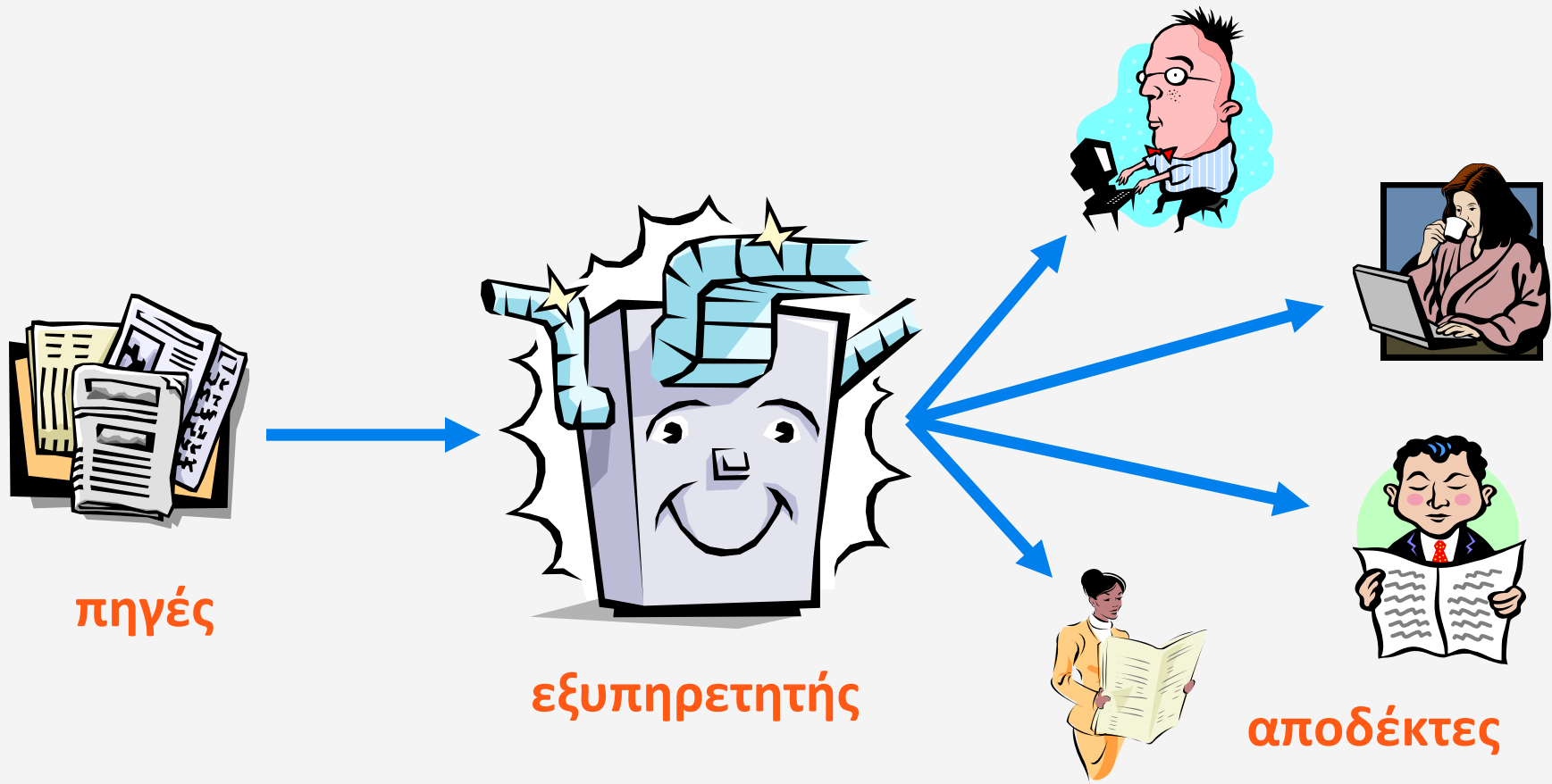
- Αναγνώριση κατηγορίας κειμένου
- Εξαγωγή πληροφορίας

Εταιρεία Αγοραστής	Αγοραζόμενη Εταιρεία	Ποσό Εξαγοράς	Ποσοστό Εξαγοράς	Χρόνος Εξαγοράς
Όμιλος Σελόντα Riopresca	Τρίτων ΑΕ Ιχθυοκαλλιέργειας	280.000 €	0,80	Παρελθόν

Τεχνολογίες: Σύντηξη πληροφορίας από πολλαπλά μέσα

A screenshot of a Windows 95/NT desktop environment. The desktop background is a light blue gradient. Several icons are visible: 'My Computer', 'My Briefcase', 'Adobe Acrobat 4.0', 'Microsoft Outlook', 'Netscape Communicator', and 'Setup for Microsoft I...'. The taskbar at the bottom shows the 'Start' button and several open application windows. The primary window is a web browser displaying a news article in Greek titled 'Μηχανή ψάχνει σημάδια ζωής στον Άρη' (Machine searches for signs of life on Mars). The article discusses the Mars Polar Lander mission and the search for water on Mars. Other open windows include 'Annotations window' with a table containing fields like 'Person', 'Location', 'Date', and 'Event', and 'Image Tagger 0.5' which is used for tagging images. A red arrow points from the 'Annotations window' to the 'Image Tagger' window. The 'Image Tagger' window shows a close-up of a Martian rock with a green circle highlighting a specific feature. The 'Annotations window' table contains the following data: Person: K.V. Chandrinou, Location: NASA archives, Date: 9-12-93, Event: Mars landing. The table also has a 'Video simulation of planet Mars' link. The 'Image Tagger' window has a 'Media time' section with a progress bar and buttons for 'Start' and 'Stop'. The 'Annotations window' has a 'File' section with a table containing the following data: Person: K.V. Chandrinou, Location: NASA archives, Date: 9-12-93, Event: Mars landing. The table also has a 'Video simulation of planet Mars' link. The 'Image Tagger' window has a 'Media time' section with a progress bar and buttons for 'Start' and 'Stop'. The 'Annotations window' has a 'File' section with a table containing the following data: Person: K.V. Chandrinou, Location: NASA archives, Date: 9-12-93, Event: Mars landing. The table also has a 'Video simulation of planet Mars' link.

Τεχνολογίες: Μοντελοποίηση Χρηστών



Αξιοποίηση πληροφορίας από κείμενα

- Ανάκτηση πληροφορίας
 - Ο χρήστης εκφράζει τα ενδιαφέροντά του χρησιμοποιώντας κάποιες λέξεις κλειδιά
 - Το σύστημα εξάγει **έγγραφα** σχετικά με τα ενδιαφέροντα του χρήστη
 - Ο χρήστης αναλύει τα έγγραφα για να εξάγει τα δεδομένα που τον ενδιαφέρουν
- Εξαγωγή πληροφορίας
 - Το σύστημα εξάγει όχι απλά τα σχετικά με τα ενδιαφέροντα του χρήστη έγγραφα, αλλά συγκεκριμένα **δεδομένα** από τα έγγραφα αυτά
 - Ο χρήστης αναλύει τα δεδομένα

Εξαγωγή πληροφορίας (ΕΠ)

- Η **εξαγωγή πληροφορίας** (information extraction/text analytics) ασχολείται με την ανάπτυξη μοντέλων και αλγορίθμων για την εύρεση «γεγονότων» και την εξαγωγή πληροφορίας για τα γεγονότα αυτά
 - Η διαδικασία ανάλυσης (πολυμεσικών) δεδομένων (κείμενα, ήχου, εικόνων, βίντεο) για την εξαγωγή περιεχομένου (ανθρώπων, τόπων, πραγμάτων, γεγονότων, προθέσεων, κ.α.) για συγκεκριμένο σκοπό (πλήρωση βάσεων δεδομένων, απαντήσεις ερωτημάτων, περιλήψεις, οπτικοποίηση, κ.α.)

Εξαγωγή πληροφορίας (2)

- Στόχοι:
 - Εύρεση ενδιαφερόντων «γεγονότων» σε έγγραφα
 - Εύρεση και κατανόηση σχετικών τμημάτων κειμένου
 - Συλλογή πληροφορίας από πολλαπλά τμήματα του κειμένου
 - Εξαγωγή σχετικής πληροφορίας για τα «γεγονότα» αυτά και αποθήκευσή της
 - Σε μια «δομημένη» αναπαράσταση: με σχέσεις (με την έννοια των σχεσιακών βάσεων), βάση γνώσης

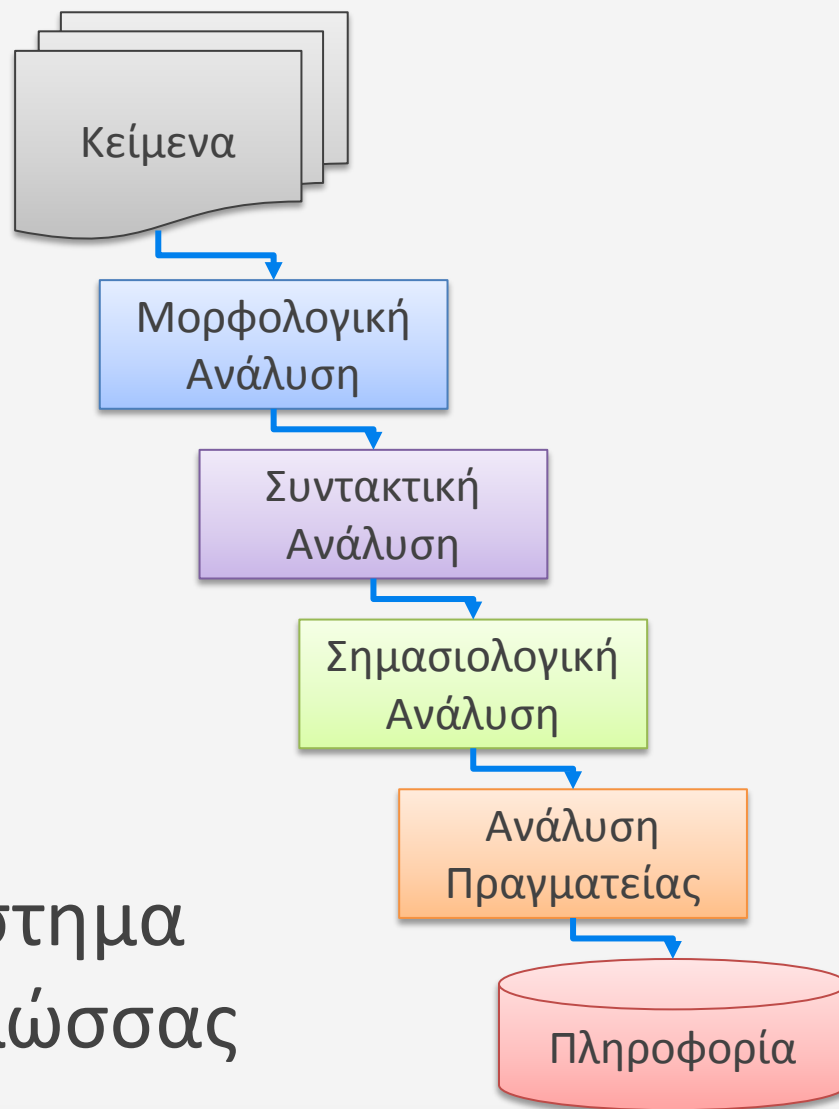
Εξαγωγή πληροφορίας (3)

- Στόχοι:
 - Οργάνωση πληροφορίας με τέτοιο τρόπο ώστε να είναι χρήσιμη στους χρήστες
 - Οργάνωση πληροφορίας σε μορφή με ακριβή σημασιολογία
 - Όστε να είναι δυνατός ο συμπερασμός (inference) από σχετικούς αλγορίθμους

Εξαγωγή πληροφορίας (4)

- Δυσκολίες:
 - Επεξεργασία αδόμητου ή ημι-δομημένου περιεχομένου
 - Αναγνώριση ονομάτων οντοτήτων, ιδιοτήτων/σχέσεων οντοτήτων, γεγονότων
 - Απαιτείται κατανόηση περιεχομένου του εγγράφου

ΕΠ από κείμενα



Ένα τυπικό σύστημα ΕΠ
είναι ουσιαστικά ένα σύστημα
επεξεργασίας φυσικής γλώσσας

ΕΠ: πρώτα στάδια

- Προ-επεξεργασία
 - Αναγνώριση λέξεων, προτάσεων, αφαίρεση HTML, κλπ.
- Μορφολογική ανάλυση
 - Αναγνώριση μερών λόγου, λημμάτων, θεμάτων, κλπ.
- Αναγνώριση χρονικών, αριθμητικών εκφράσεων κλπ.
 - και μετατροπή τους σε κανονικές μορφές
- Αναγνώριση ονομάτων οντοτήτων
 - Ονόματα προσώπων, οργανισμών, εταιριών, κλπ.
 - Ταίριασμα διαφορετικών μορφών
 - «ΟΤΕ» και «Οργανισμών Τηλεπικοινωνιών Ελλάδος»

ΕΠ: επόμενα στάδια

- Επίλυση αναφορικών εκφράσεων
 - Π.χ. «ο οργανισμός ανακοίνωσε απολύσεις...»
- Εντοπισμός σχέσεων μεταξύ ονομάτων οντοτήτων
 - Και κατηγοριοποίηση σε κατάλληλες κατηγορίες
- Εντοπισμός «γεγονότων»
 - Συμβάντα που περιγράφονται στα κείμενα
 - Χρόνοι γεγονότων
 - Συσχετισμός εντοπισμένων σχέσεων με ρόλους των γεγονότων
- Εντοπισμός σχέσεων μεταξύ γεγονότων

Διεθνή Συνέδρια Αξιολόγησης

- Message Understanding Conferences (MUC)
 - 1987 – 1998
- Document Understanding Conferences (DUC)
 - 2001 – 2007
- Text Analysis Conferences (TAC)
 - 2008 – 2012

Συνέδρια Αξιολόγησης MUC

- MUCK Μηνύματα Ναυτικού (1987)
- MUCK-II Μηνύματα Ναυτικού (1989)
- MUC-3 Ειδήσεις για τρομοκρατικές επιθέσεις (1991)
- MUC-4 Ειδήσεις για τρομοκρατικές επιθέσεις (1992)
- MUC-5 Επιχειρηματικές ειδήσεις (joint ventures, micro-electronics products) (1993)
- MUC-6 Επιχειρηματικές ειδήσεις (management succession) (1995)
- MUC-7 Ειδήσεις για εκτοξεύσεις πυραύλων (1998)

Εργασίες αξιολόγησης

- Αναγνώριση ονομάτων οντοτήτων (Named Entity Recognition)
- Προσδιορισμός κοινών αναφορών οντοτήτων (Co-reference Identification)
- Εξαγωγή πληροφορίας για τις οντότητες (Template Elements Filling)
- Εξαγωγή πληροφορίας για τις σχέσεις μεταξύ των οντοτήτων (Template Relations Filling)
- Εξαγωγή πληροφορίας για τα γεγονότα στα οποία εμπλέκονται οι οντότητες (Scenario-Based Template Elements Filling)

Αναγνώριση ονομάτων οντοτήτων

Προσδιορισμός και κατηγοριοποίηση ονοματικών εκφράσεων που αναφέρονται σε συγκεκριμένες οντότητες στο κείμενο

- Παραδείγματα εκφράσεων
 - ονόματα οργανισμών
 - ονόματα προσώπων
 - ονόματα τοποθεσιών
 - χρονικές εκφράσεις (ημερομηνία, ώρα, ...)
 - αριθμητικές εκφράσεις
 - ...

Προσδιορισμός κοινών αναφορών οντοτήτων

Προσδιορισμός φράσεων στο κείμενο που αναφέρονται στην ίδια οντότητα

- Παραδείγματα εκφράσεων
 - αντωνυμίες (αυτός, αυτή, εκείνος, ...)
 - οριστικά, αόριστα άρθρα (μία εταιρεία, η εταιρεία Χ, ...)
 - προσδιορισμοί (ο Πρόεδρος, ο υπ. Ανάπτυξης)
 - διαφορετικές εκφράσεις ονομάτων οντοτήτων (Γιάννης Παπαδόπουλος, Παπαδόπουλος, Γ. Παπαδόπουλος, ...)
 - ...

Εξαγωγή πληροφορίας για τις οντότητες

Προσδιορισμός φράσεων στο κείμενο που προσδιορίζουν τις οντότητες

- Παραδείγματα

<ORGANIZATION> :=

ORG_NAME:

ORG_DESCRIPTOR:

ORG_TYPE:

ORG_ALIAS:

ORG_COUNTRY:

<PERSON> :=

PER_NAME:

PER_TITLE:

PER_ALIAS:

Εξαγωγή πληροφορίας για τις σχέσεις μεταξύ των οντοτήτων

Προσδιορισμός σχέσεων μεταξύ των οντοτήτων στο συγκεκριμένο γεγονός, για το οποίο θέλουμε να εξάγουμε πληροφορία

- Παραδείγματα

- εξαγορές επιχειρήσεων

- σχέση αγοραστή-αγοραζόμενου μεταξύ του οργανισμού-1 και του οργανισμού-2

- παραιτήσεις στελεχών επιχειρήσεων

- σχέση προσώπου με οργανισμό

- παραγγελίες τμημάτων αεροπλάνων

- σχέση οργανισμού με αντικείμενο

- ...

Εξαγωγή πληροφορίας για το γεγονός στο οποίο εμπλέκονται οι οντότητες

Προσδιορισμός εκφράσεων με επιπλέον πληροφορίες (πέρα από τις σχέσεις μεταξύ οντοτήτων) στο συγκεκριμένο γεγονός

- Παραδείγματα

- χρόνος του γεγονότος (παρελθόν, παρόν, μέλλον)
- τόπος του γεγονότος
- σε ένα γεγονός παραίτησης, η θέση εργασίας
- σε ένα γεγονός εξαγοράς, το κεφάλαιο της εξαγοράς
- ...

Παράδειγμα: Κείμενο από τη Wall Street Journal

```
<DOC>
<DOCID> wsj93_019.0225 </DOCID>
<DOCNO> 930512-0128. </DOCNO>
<HL>    Who's News:
@  R.G. Barry Corp. </HL>
<DD> 05/12/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B8 </SO>
<CO>    RGB </CO>
<IN> FOOTWEAR MANUFACTURERS (FOT), TEXTILES (TEX) </IN>
<TXT>
<p>
    R.G. BARRY Corp. (Pickerington, Ohio) -- Lee Abraham, 65 years
old, former chairman and chief executive officer of Associated
Merchandising Corp., New York, was named to the board of the
footwear manufacturer. He succeeds Allan J. Bloostein, 63, who
retired from the board for personal reasons. The size of the board
remains at nine members.
</p>
</TXT>
</DOC>
```

Παράδειγμα: κενή εγγραφή για γεγονότα

- Που αφορούν αλλαγές στο προσωπικό επιχειρήσεων (management succession events)

```
<TEMPLATE> :=  
    DOC_NR:.....  
    CONTENT:.....  
<SUCCESSION_EVENT> :=  
    ORGANIZATION: .....  
    POST: .....  
    IN_AND_OUT:.....  
    VACANCY_REASON:.....  
<IN_AND_OUT> :=  
    IO_PERSON: .....  
    NEW_STATUS: .....  
    ON_THE_JOB: .....  
<ORGANIZATION> :=  
    ORG_NAME: .....  
    ORG_DESCRIPTOR: .....  
    ORG_TYPE: .....  
    ORG_LOCALE: .....  
    ORG_COUNTRY: .....  
    ORG_ALIAS: .....  
<PERSON-1.txt.cl-8> :=  
    PER_NAME: .....  
    PER_ALIAS: .....  
    PER_TITLE: .....
```

Παράδειγμα: εγγραφή με πληροφορία

<TEMPLATE-1.txt.cl-1> :=

DOC_NR: "1.txt.cl"

CONTENT: <SUCCESSION_EVENT-1.txt.cl-12>

<SUCCESSION_EVENT-1.txt.cl-12> :=

SUCCESSION_ORG: <ORGANIZATION-1.txt.cl-15>

POST: "chairman"

IN_AND_OUT: <IN_AND_OUT-1.txt.cl-4>

VACANCY_REASON: REASSIGNMENT

<IN_AND_OUT-1.txt.cl-4> :=

IO_PERSON: <PERSON-1.txt.cl-8>

NEW_STATUS: OUT

ON_THE_JOB: NO

<ORGANIZATION-1.txt.cl-15> :=

ORG_NAME: "Associated Merchandising Corp"

ORG_TYPE: COMPANY

ORG_LOCALE: New York CITY

ORG_COUNTRY: United States

<PERSON-1.txt.cl-8> :=

PER_NAME: "Lee Abraham"

Μέθοδοι αξιολόγησης συστημάτων (1)

- Αναλυτές εξάγουν τη “σωστή” πληροφορία από αντιπροσωπευτικά κείμενα της θεματικής περιοχής, αποθηκεύοντάς την στις εγγραφές της βάσης
- Τα συστήματα ΕΠ εξάγουν πληροφορία από τα ίδια κείμενα, αποθηκεύοντάς την

Μέθοδοι αξιολόγησης συστημάτων (2)

- Τα αποτελέσματα των συστημάτων συγκρίνονται με τα αποτελέσματα των αναλυτών
- Η σύγκριση γίνεται ως προς δύο παραμέτρους:

$$\text{Recall} = \frac{\text{Σωστές απαντήσεις συστήματος}}{\text{Συνολικές σωστές απαντήσεις}}$$

$$\text{Precision} = \frac{\text{Σωστές απαντήσεις συστήματος}}{\text{Συνολικές απαντήσεις συστήματος}}$$

Confusion Matrix (1)

- Επίσης: contingency table, error matrix

		Predicted (Απαντήσεις Συστήματος)	
Actual (Σωστά)		Positive	Negative
	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$$

Confusion Matrix (2)

- Παράδειγμα (http://en.wikipedia.org/wiki/Confusion_matrix)

		Predicted (Απαντήσεις Συστήματος)		
Actual (Σωστά)		Γάτα	Σκύλος	Λαγός
	Γάτα	5	3	0
	Σκύλος	2	3	1
	Λαγός	0	2	11

5 True Positives
(Σωστές γάτες που κατηγοριοποιήθηκαν σωστά σαν γάτες)

3 False Negatives
(Γάτες που κατηγοριοποιήθηκαν λανθασμένα σαν σκύλοι)

2 False Positives
(Σκύλοι που κατηγοριοποιήθηκαν λανθασμένα σαν γάτες)

17 True Negatives
(Όλα τα υπόλοιπα ζώα που Κατηγοριοποιήθηκαν σωστά σαν μη γάτες)

Confusion Matrix (3)

Definition (classification context) [edit]

For classification tasks, the terms **true positives**, **true negatives**, **false positives**, and **false negatives** (see also **Type I and type II errors**) compare the results of the classifier under test with trusted external judgments. The terms *positive* and *negative* refer to the classifier's prediction (sometimes known as the *expectation*), and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment (sometimes known as the *observation*).

Let us define an experiment from **P** positive instances and **N** negative instances for some condition. The four outcomes can be formulated in a 2x2 *contingency table* or *confusion matrix*, as follows:

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	True positive	False positive (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	False negative (Type II error)	True negative	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$

Precision and recall are then defined as:^[5]

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Recall in this context is also referred to as the true positive rate or **sensitivity**, and precision is also referred to as **positive predictive value** (PPV); other related measures used in classification include true negative rate and **accuracy**.^[6] True negative rate is also called **specificity**.

$$\text{True negative rate} = \frac{tn}{tn + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Terminology and derivations from a confusion matrix

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews correlation coefficient (MCC)

$$TP \times TN - FP \times FN$$

$$\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Source: Fawcett (2006).^[4]

http://en.wikipedia.org/wiki/Precision_and_recall

Μια συνδυαστική μετρική: F

- Μια μετρική που συνδυάζει ανάκληση/ακρίβεια:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$

- Harmonic mean: F_1

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

- F_2 : μεγαλύτερο βάρος στην ανάκληση
- $F_{0.5}$: μεγαλύτερο βάρος στην ακρίβεια

Αξιολόγηση συστημάτων ΕΠ

- Στο MUC-7 (για την Αγγλική γλώσσα)

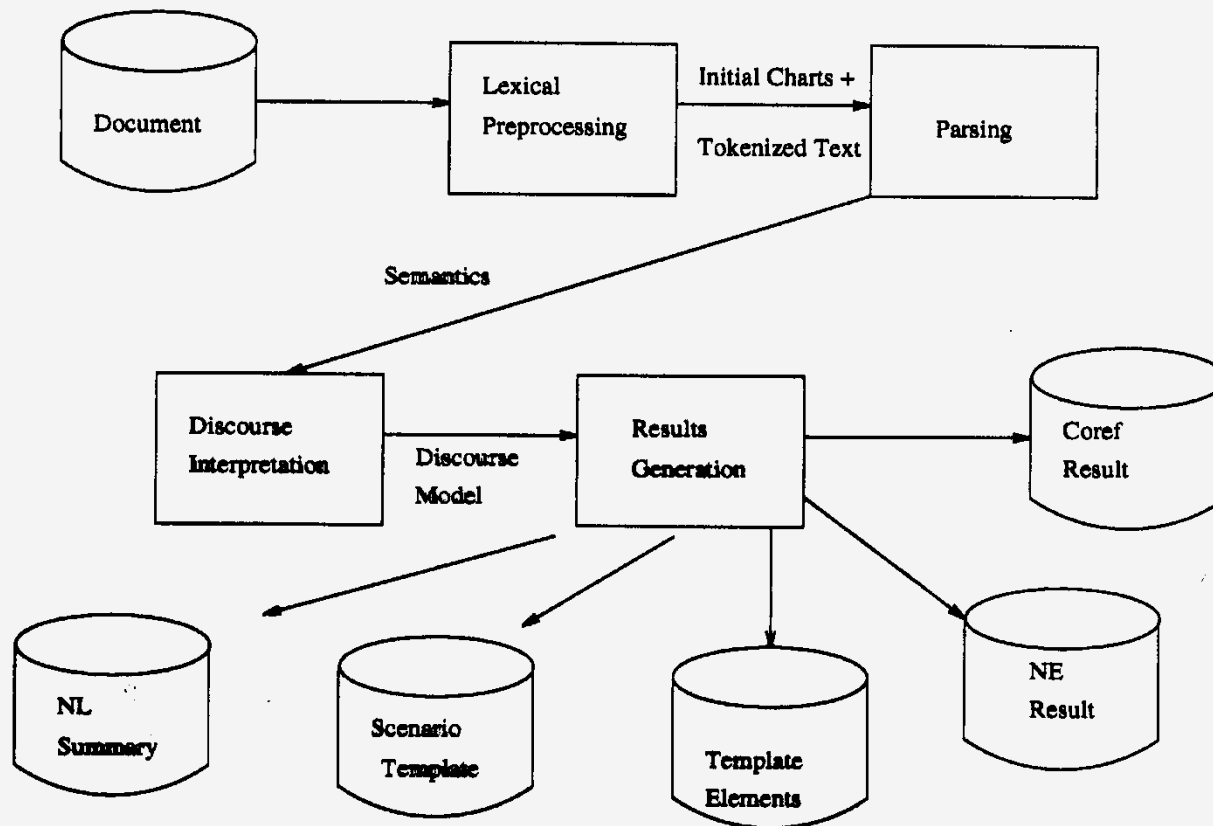
Task	Recall (%)	Precision (%)
Named Entity (Edinburgh)	92	95
Co-reference (Sheffield)	56	69
Template Element (SRA)	86	87
Scenario Template (SRA)	42	65

Συμπεράσματα αξιολόγησης MUC

- Παρά τα πολύ καλά αποτελέσματα σε επιμέρους εργασίες και εφαρμογές, συνολικά οι επιδόσεις των συστημάτων κυμαίνονται στο 60 – 65 %
- Όταν η ΕΠ γίνεται από ανθρώπους (annotators), αυτοί συμφωνούν (interannotator agreement) σε ποσοστό 60 – 80 %
 - κάτι που δείχνει και τη δυσκολία της εργασίας της ΕΠ

Ένα τυπικό σύστημα ΕΠ

- LaSIE (Large Scale IE), University of Sheffield



Προσαρμογή (1)

- Σε νέες Θεματικές Περιοχές και Γλώσσες
 - Εξαγωγή Πληροφορίας: μία από τις πλέον απαιτητικές διαδικασίες επεξεργασίας φυσικής γλώσσας
 - Προβλήματα: εντοπισμός σημαντικών γεγονότων και οντοτήτων στο κείμενο, συσχέτιση των οντοτήτων με τα γεγονότα, ...
 - Γενική επίλυση τέτοιων προβλημάτων: μακρινός στόχος για τα σημερινές «μηχανές» επεξεργασίας φυσικής γλώσσας

Προσαρμογή (2)

- Πρακτική λύση: εστίαση του συστήματος σε μια συγκεκριμένη θεματική περιοχή/γλώσσα (π.χ. συγχωνεύσεις οργανισμών για κείμενα της Αγγλικής)
 - επιτρέποντας τη χρήση γλωσσικών πόρων για τη συγκεκριμένη θεματική περιοχή/γλώσσα (π.χ. λίστα ονομάτων γνωστών ελληνικών οργανισμών)
 - Ο περιορισμός του προβλήματος σε μια συγκεκριμένη θεματική περιοχή/γλώσσα κάνει το πρόβλημα πρακτικά επιλύσιμο
 - αλλά ταυτόχρονα δημιουργεί την ανάγκη για συνεχή προσαρμογή του συστήματος ΕΠ σε νέες θεματικές περιοχές/γλώσσες
 - Δαπανηρή η προσαρμογή εξαιτίας των γλωσσικών πόρων που απαιτούνται

Προσεγγίσεις προσαρμογής

- Δύο προσεγγίσεις:
 - με χρήση κανόνων γραμμένων από «ειδικούς»
(knowledge engineering approach)
 - με αυτόματη εκπαίδευση
(automatic training approach)

Πολυγλωσσική (Multilingual) ΕΠ

- Δύο προσεγγίσεις:
 - Σύστημα ΕΠ, που εκτελεί μονογλωσσική ΕΠ, αλλά για περισσότερες από μία γλώσσες (διαφορετική έκδοση του συστήματος για κάθε γλώσσα):
 - Αρχική (source) και τελική (extraction) γλώσσα: ίδιες
 - Σύστημα ΕΠ που εκτελεί διαγλωσσική (cross-lingual) ΕΠ:
 - Αρχική (source) και τελική (target) γλώσσα: διαφορετικές

Ανάπτυξη εφαρμογών ΕΠ (1)

- Χρήση χειρωνακτικά κατασκευασμένων κανόνων
 - Μελέτη σώματος κειμένων αντιπροσωπευτικών της θεματικής περιοχής (training corpus)
 - Χειρωνακτική κατασκευή των γραμματικών από ειδικούς (knowledge engineers)
 - Επαναληπτική διαδικασία για τη βελτίωση των επιδόσεων του συστήματος

Ανάπτυξη εφαρμογών ΕΠ (2)

- Αυτόματη εκπαίδευση
 - Χρήση τεχνικών μηχανικής μάθησης
 - Επισημείωση σώματος αντιπροσωπευτικών κειμένων (annotation of training corpus)
 - Αυτόματη εκμάθηση κανόνων από τα επισημειωμένα κείμενα
 - Αυτόματη εκμάθηση κανόνων από αλληλεπίδραση με το χρήστη

Ανάπτυξη εφ. ΕΠ: χρήση κανόνων (1)

- Πλεονεκτήματα
 - Η συγγραφή των κανόνων από έμπειρους «knowledge engineers» οδηγεί στη δημιουργία συστημάτων με καλές επιδόσεις
 - Σχετικά εύκολη προσαρμογή σε τυχόν αλλαγή των προδιαγραφών
 - Τα καλύτερα συστήματα ΕΠ βασίζονται σε χειρωνακτικά κατασκευασμένους κανόνες

Ανάπτυξη εφ. ΕΠ: χρήση κανόνων (2)

- Μειονεκτήματα
 - Χρονοβόρα διαδικασία
 - Προϋποθέτει την ύπαρξη των κατάλληλων γλωσσικών πόρων
 - Δύσκολη η εύρεση των ειδικών (knowledge engineers)

Ανάπτυξη εφ. ΕΠ: αυτόματη εκπαιδ. (1)

- Πλεονεκτήματα
 - Διευκολύνει την προσαρμογή σε νέες θεματικές περιοχές
 - Δεν απαιτεί ειδικούς knowledge engineers
 - Με χρήση του κατάλληλου σώματος εκπαίδευσης οδηγεί στη δημιουργία συστημάτων με καλές επιδόσεις

Ανάπτυξη εφ. ΕΠ: αυτόματη εκπαιδ. (2)

- Μειονεκτήματα
 - Δύσκολη η δημιουργία του κατάλληλου σώματος εκπαίδευσης (απαιτείται συνήθως επισημείωση μεγάλου όγκου δεδομένων)
 - Τυχόν αλλαγή των προδιαγραφών μπορεί να απαιτήσει επανασημείωση του σώματος εκπαίδευσης για την επανεκπαίδευση του συστήματος

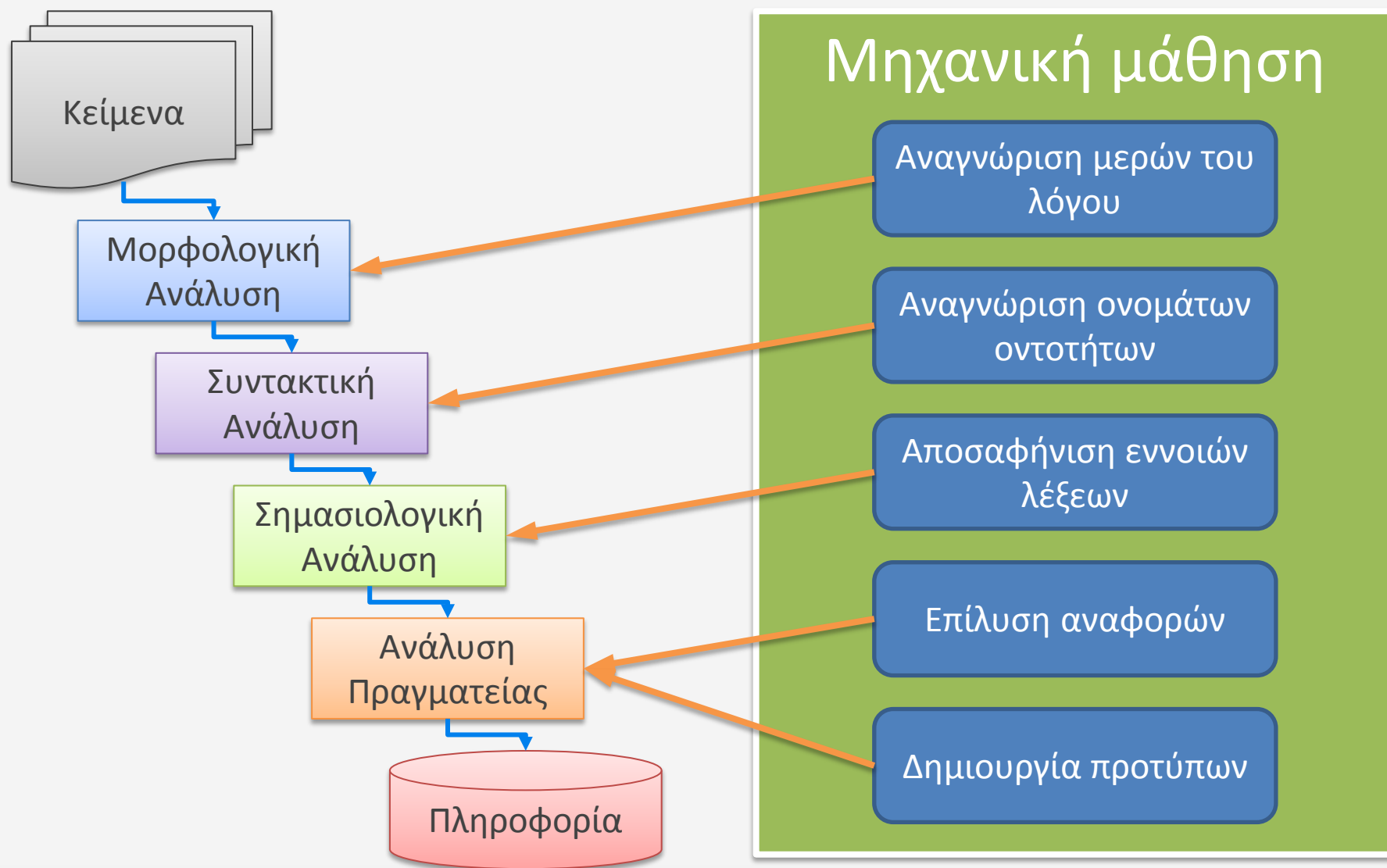
Πότε μια προσέγγιση είναι κατάλληλη; (1)

- Χρήση κανόνων όταν:
 - υπάρχουν οι απαραίτητοι γλωσσικοί πόροι (π.χ. λεξικά, λίστες) και εργαλεία που υποστηρίζουν τη συγγραφή κανόνων (π.χ. pattern editors)
 - οι προδιαγραφές είναι πιθανό να αλλάξουν
 - υπάρχουν οι «ειδικοί» για τη συγγραφή κανόνων
 - ο όγκος ορισμένων από τα δεδομένα εκπαίδευσης είναι σχετικά μικρός
 - είναι ιδιαίτερα σημαντικό το τελικό σύστημα να έχει όσο γίνεται καλύτερη απόδοση

Πότε μια προσέγγιση είναι κατάλληλη; (2)

- Χρήση αυτόματης εκπαίδευσης όταν:
 - υπάρχει το κατάλληλο σώμα εκπαίδευσης ή είναι σχετικά εύκολη η δημιουργία του (κυρίως όσον αφορά την επισημείωση και τον όγκο των δεδομένων εκπαίδευσης)
 - οι προδιαγραφές δεν είναι πιθανό να αλλάξουν
 - δεν υπάρχουν οι «ειδικοί» για τη συγγραφή κανόνων
 - είναι αρκετό το τελικό σύστημα να έχει απλά ικανοποιητική απόδοση

Ανάπτυξη Εφαρμογών ΕΠ



Αναγνώριση ονομάτων οντοτήτων

- Συχνά χρησιμοποιούνται ετικέτες B-I-O
 - Για κάθε οντότητα
 - B: αρχή οντότητας (1^η λέξη) (begin)
 - I: επόμενες λέξεις που απαρτίζουν την οντότητα (inside)
 - O: λέξεις που δεν αποτελούν μέρος οντότητας (other)
- Το πρόβλημα μετατρέπεται σε ακολουθιακή κατηγοριοποίηση (sequence labeling)
- Μπορούν να χρησιμοποιηθούν αλγόριθμοι επιβλεπόμενης μάθησης
 - Π.χ. HMM, ME, SVM, CRF, κλπ.

Παράδειγμα IOB

[ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PERS Tim Wagner] said.

Words	Label
American	B _{ORG}
Airlines	I _{ORG}
,	O
a	O
unit	O
of	O
AMR	B _{ORG}
Corp.	I _{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B _{PERS}
Wagner	I _{PERS}
said	O
.	O

Αναγνώριση ονομάτων οντοτήτων

- Συχνά το ποιες ιδιότητες/χαρακτηριστικά χρησιμοποιούνται έχει μεγαλύτερη επίπτωση από την επιλογή του αλγορίθμου μηχανικής μάθησης
- Πιθανές ιδιότητες:
 - Μορφή (shape) λεκτικής μονάδας
 - Αν περιλαμβάνει κεφαλαίους/πεζούς χαρακτήρες, σύμβολα, κλπ.
 - Θέμα, λήμμα, κατάληξη, κλπ.
 - Μέρος του λόγου
 - Κατηγορία από ειδικές λίστες (π.χ. «Α.Ε.», «Γιώργος»)
 - N-γράμματα, κοντινές λέξεις, κλπ.