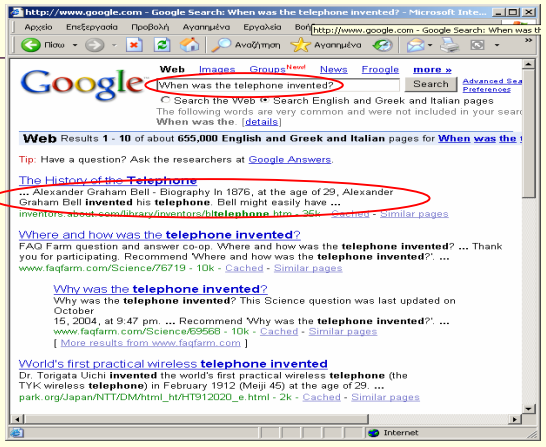
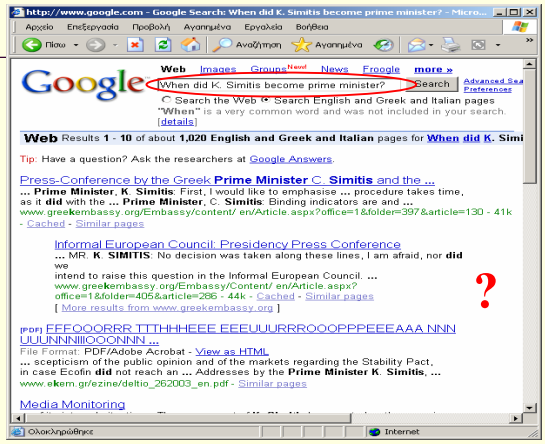
Χειρισμός ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων για συλλογές εγγράφων

Ι. Ανδρουτσόπουλος, Δ. Γαλάνης, Ι. Μηλιαράκη

Τμήμα Πληροφορικής
Οικονομικό Πανεπιστήμιο Αθηνών





Συστήματα ερωταποκρίσεων

- Υποτομέας της **επεξεργασίας φυσικής γλώσσας**.
  - Φιλοδοξεί να βελτιώσει τις σημερινές μηχανές αναζήτησης.
  - Μεγάλη προϊστορία σε συστήματα ερωταποκρίσεων για ΒΔ.
- Εδώ εύρεση σε μια **συλλογή εγγράφων** (ή τον Ιστό) απαντήσεων σε **ερωτήσεις** των χρηστών.
  - *"Who invented the telephone?"*, *"How much did Mercury spend on advertising in 1993?"*.
- Αργότερα θα εστιαστούμε στις **ερωτήσεις ορισμού**.
  - *"What is a tsunami?"*, *"What are pathogens?"*, *"Who was Duke Ellington?"*.
- Οι απαντήσεις συνήθως είναι **αποσπάσματα** εγγράφων (π.χ. 250 χαρακτήρων) ή **ακριβείς απαντήσεις** (π.χ. ονόματα).
  - *"When Graham Bell invented the telephone, some people thought..."*, *"Graham Bell"*.

Εκτενέστερη γλωσσική επεξεργασία

αντικείμενο

> When was the ***telephone invented?***

Χρονική ερώτηση. Απαιτεί χρονικό προσδιορισμό.

- The ***telephone*** was ***invented*** by Alexander Graham Bell in 1876.

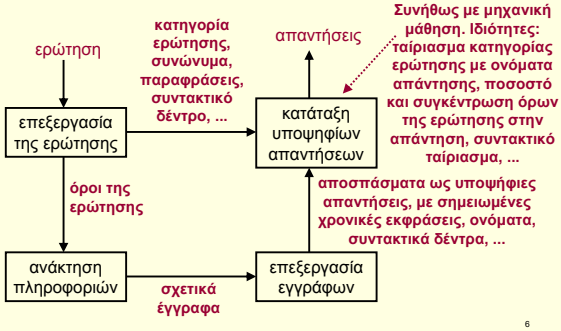
αντικείμενο

- The ***telephone*** was ***invented*** by Graham Bell.
- He ***invented*** the ***telephone*** on March 10, 1876.
- Fifteen years later, Alexander Graham Bell ***invented*** the ***telephone*** and became multi-millionaire.

αντικείμενο

- In 1992 Dr. Johnson ***invented*** a new device that works with an ordinary ***telephone*** line.

Συστήματα ερωταποκρίσεων



Ερωτήσεις ορισμού

- **Πολύ συχνές**. 27% στο QA track του TREC-2001.
- **Δύσκολες** για τα τυπικά συστήματα ερωτ/σεων.
  - Η κατηγορία της ερώτησης **δεν δημιουργεί προσδοκίες** συγκεκριμένων κατηγοριών **ονομάτων** στην απάντηση.
    - *"Who is the president of Greece?"* → πρόσωπο
    - *"What is a tsunami?"* → ?
  - Πολύ **λίγοι όροι** στις ερωτήσεις, πάντα μικρό ποσοστό όρων της ερώτησης στις υποψήφιες απαντήσεις.
- Μεγάλη **ποικιλία εκφράσεων** σηματοδοτεί ορισμούς αλλά όχι πάντα:
  - *"... the giant wave known as tsunami..."*
  - *"... the fear of creating hazardous microorganisms, or pathogens, is overstated."*

Διάρθρωση της ομιλίας

- Εισαγωγή:
  - Συστήματα ερωταποκρίσεων
  - Ερωτήσεις ορισμού.
- 1η ενότητα (πτυχιακή Μηλιαράκη, άρθρο COLING):
  - Προηγούμενες μέθοδοι για ερωτήσεις ορισμού.
  - Επιβλεπόμενη μέθοδος μάθησης.
  - Πειράματα με άρθρα εφημερίδων από το TREC.
- 2η ενότητα (πτυχιακή Γαλάνη, υπό δημοσίευση):
  - Πειράματα με ιστοσελίδες.
  - Μη επιβλεπόμενη μορφή της μεθόδου.

1η ενότητα: το ζητούμενο

- Είσοδος: ένας **όρος** προς ορισμό (πιθανώς πολλών λέξεων) και τα **κορυφαία έγγραφα** που επέστρεψε μια μηχανή ανάκτησης πληροφοριών για τον όρο.
  - Υποθέσεις: Διατίθεται ένας ταξινομητής που ξεχωρίζει τις ερωτήσεις ορισμού (βλ. πτυχιακή Μαυροειδή) και μια μονάδα που εντοπίζει τον **όρο-στόχο** μέσα στην ερώτηση.
- Έξοδος: το πολύ **5 αποσπάσματα** των 250 χαρακτήρων το καθένα.
- Ορθή έξοδος αν τουλάχιστον **ένα από τα αποσπάσματα** περιέχει αποδεκτό ορισμό του όρου.
  - Χρησιμοποιούμε δεδομένα των TREC-2000 και 2001.
  - Τα αποσπάσματα κρίνονται με τα **πρότυπα** (Perl patterns) του TREC.
  - **Ορισμοί ενός αποσπάσματος**, όχι όπως στο TREC-2003.

Παραδείγματα αποσπασμάτων

Για τον όρο «pathogens»:

✓ …considerations as nutrient availability. In particular, the panel concluded that the fear of creating hazardous microorganisms, or pathogens, is overstated. "It is highly unlikely that moving one or a few genes from a pathogen to…

✘ …definite intraspecial physiological and morphological diversity. Ph. helianthi thrives at higher temperatures than other sunflower pathogens (Sclerotinia sclerotiorum and Botrytis cinerea) do. In various nutrient media, Ph. helianthi …

Παραδείγματα αποσπασμάτων – II

Για τον όρο «Moulin Rouge»:

✓ …among the guests at a benefit gala that celebrated the centennial of the Moulin Rouge, a Paris landmark. The Moulin Rouge, a cabaret under a giant red windmill, is known worldwide, especially from the posters of the Henri de Toulouse-Lautrec …

✘ …throughout was the foundation's president, Danielle Mitterrand, wife of the French president. "Welcome to the Moulin Rouge, the heart and soul of Paris", said comedian Jerry Lewis, who wisecracked his way through the introduction…

Prager et al. (2001, 2002)

- Ένας ορισμός συχνά περιέχει ένα **υπερώνυμο** του όρου-στόχου.
  - *An amphibian is an animal that lives both in water...*
  - *"amphibian" < "carnivore" < ... < "animal" < ...*
  - Το υπερώνυμο μόνο του ίσως είναι φτωχή απάντηση.
- Βρες τα **«καλύτερα» υπερώνυμα**.
  - Εμφανίζονται συχνά μαζί με τον όρο-στόχο στα κείμενα της συλλογής και δεν βρίσκονται πολύ πιο ψηλά από τον όρο-στόχο στην ιεραρχία του WordNet.
- Βρες και **κατάταξε** (π.χ. με κεντροειδές) τα **αποσπάσματα** (προτάσεις) που περιέχουν τον **όρο-στόχο** και ένα από τα **καλύτερα υπερώνυμά** του.
- Προβλήματα: Ο όρος-στόχος ίσως δεν υπάρχει στο WordNet. Τα υπερώνυμα συχνά δεν βοηθούν. Τι γίνεται με γλώσσες για τις οποίες δεν υπάρχει WordNet;

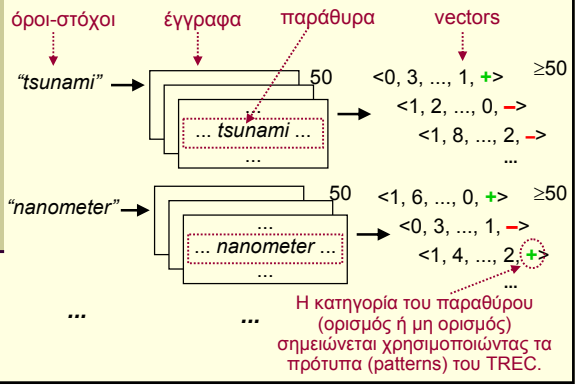
Joho et al. (2000, 2001)

- **9 χειρωνακτικά κατασκευασμένα πρότυπα** (patterns), μερικά από τη Hearst (1998).
  - *"[στόχος] (and | or) other"*, *"[στόχος], (a | an | the)"*
  - Η ακρίβεια (precision) των προτύπων μετριέται σε ένα σώμα εκπαίδευσης.
- Εντόπισε τις **προτάσεις** που περιέχουν τον όρο-στόχο και κατάταξέ τις βάσει **3 ιδιοτήτων**:
  - *KPW*: **ακρίβεια** προτύπου που ταίριαξε, αν υπάρχει.
  - *SN*: **τακτικός αριθμός** της πρότασης στο έγγραφό της, μεταξύ προτάσεων που περιέχουν τον όρο-στόχο.
  - *WC*: τι **ποσοστό των λέξεων** που είναι πολύ **συχνές** (20 συχνότερες) μεταξύ όλων των υποψηφίων απαντήσεων (προτάσεων) εμφανίζεται στην πρόταση.
- Κατάταξε τις προτάσεις με το **ζυγισμένο άθροισμα** των 3 ιδιοτήτων και χειρωνακτική ρύθμιση βαρών.
- Προβλήματα: περιορισμένα πρότυπα, ρύθμιση βαρών.

Η δική μας προσέγγιση

- **Συνδυασμός ιδεών** από προηγούμενες προσεγγίσεις με τεχνικές **μηχανικής μάθησης**.
- Πολύ **περισσότερα πρότυπα**, που παράγονται **αυτόματα**.
- **Baseline**: επανυλοποίηση της μεθόδου των Prager et al. (με αποσπάσματα 250 χαρακτήρων).
  - Χωρίς τη συνάρτηση κατάταξης. Αν πολλά αποσπάσματα περιέχουν τον όρο-στόχο και «καλύτερα» υπερώνυμα, χρησιμοποιούμε την κατάταξη των αντιστοίχων εγγράφων (RK) που επέστρεψε η μηχανή ανάκτησης πληροφοριών.
- **4 διαμορφώσεις μεθόδου μηχανικής μάθησης**.
  - Όλες με SVM, απλό γραμμικό πυρήνα.
  - Διαφορετικές ιδιότητες σε κάθε διαμόρφωση.

Διαμορφώσεις SVM – εκπαίδευση



Διαμορφώσεις SVM – συνέχεια

- Εκπαίδευση: το SVM μαθαίνει να κατατάσσει διανύσματα (παράθυρα) ως ορισμούς ή μη ορισμούς.
- Κατά τη χρήση, δοθέντος ενός όρου-στόχου:
  - Πάρε από τη μηχανή ανάκτηση πληροφοριών τα 50 **κορυφαία έγγραφα**.
  - Συγκέντρωσε τα **παράθυρα** των εγγράφων, μετάτρεψέ τα σε **διανύσματα**, κατάταξέ τα με το SVM.
  - Επίστρεψε τα **5 παράθυρα** που αντιστοιχούν στα διανύσματα για τα οποία το SVM ήταν **περισσότερο σίγουρο** ότι αποτελούν ορισμούς.
- **10-πλή διασταυρωμένη επικύρωση** σε όλα τα πειράματα αυτής της ενότητας.

Διαμόρφωση 1: προσομοίωση Joho et al.

- **Ιδιότητες**:
  - Τα ***SN*** (τακτικός αριθμός) και ***WC*** (% συχνών λέξεων) των Joho et al.
  - Μια **δυαδική ιδιότητα** για κάθε **πρότυπο** των Joho et al. και 4 επιπλέον πρότυπα.
  - ***RK*** (η κατάταξη των εγγράφων της μηχανής ανάκτησης πληροφοριών).
- Κατά μία έννοια **προσομοίωση** της μεθόδου των **Joho et al.** αλλά καλύτερη γιατί:
  - χρησιμοποιούμε μερικές **επιπλέον ιδιότητες**
  - χρησιμοποιούμε **SVM** αντί για ζυγισμένο άθροισμα με χειρωνακτικά ρυθμισμένα βάρη.

Διαμόρφωση 2: προσθήκη WordNet

- Όπως η διαμόρφωση 1 αλλά με **επιπλέον ιδιότητα**:
  - Δείχνει αν το παράθυρο περιέχει ένα από τα **«καλύτερα» υπερώνυμα** του όρου-στόχου.
  - Όπως επιστρέφονται από την επανυλοποίηση της μεθόδου των Prager et al.
- Κατά μία έννοια **συνδυασμός** των μεθόδων των Prager et al. και Joho et al.
  - Οι μέθοδοι συνεισφέρουν ιδιότητες.
  - Το SVM αποφασίζει τι βάρος θα δώσει σε κάθε ιδιότητα.
- Πιθανό πρόβλημα: Ίσως η ιδιότητα της μεθόδου Prager et al. (WordNet) να «χάνεται» μεταξύ των περισσότερων ιδιοτήτων της μεθόδου των Joho et al.
  - Εναλλακτική προσέγγιση: δύο μόνο ιδιότητες, που θα έδειχναν τις ετυμηγορίες των δύο μεθόδων (stacking).

Διαμόρφωση 3: προσθήκη *n*-γραμμάτων

- Όπως η διαμόρφωση 2 αλλά με **m πρόσθετες δυαδικές ιδιότητες** που αντιστοιχούν σε **αυτόματα αποκτηθέντα πρότυπα** (patterns).
  - Στα πειράματά μας $100 \leq m \leq 300$.
- **Υποψήφια πρότυπα**:
  - Όλα τα ***n*-γράμματα λεκτικών μονάδων** ($1 \leq n \leq 3$) που εμφανίζονται αμέσως **πριν ή μετά τον όρο-στόχο** στα κείμενα εκπαίδευσης.
  - π.χ. *"[στόχος] , which is"* , *". A [στόχος]"* .
  - **Κατώφλι**: ≥10 εμφανίσεις για κάθε υποψήφιο πρότυπο.
  - Τα υποψήφια πρότυπα **ταξινομούνται κατά ακρίβεια**.
- Κρατούμε τα ***m* κορυφαία πρότυπα** της ταξινόμησης.

Αποκτηθέντα πρότυπα

- **Ξανα-ανακάλυψε** αρκετά πρότυπα της διαμόρφωσης 1 και εύλογες παραλλαγές.
  - *"[στόχος] is one"*, *", (a | an | the) [στόχος]"* .
- Μερικά πρότυπα φαίνονται **περίεργα**, αλλά αποδεικνύονται **εύλογα** μετά από μελέτη. **Δύσκολο** να κατασκευαστούν χειρωνακτικά.
  - π.χ. *". [στόχος]"* , *". An [στόχος]"* .
- Μερικά πρότυπα ήταν για **συγκεκριμένες θεματικές** περιοχές.
  - Πολλά πρότυπα π.χ. για ασθένειες, λόγω του μεγάλου αριθμού άρθρων/ερωτήσεων για ιατρικά θέματα στη συλλογή του TREC.
  - Δείχνει ότι η μέθοδος μπορεί να προσαρμοστεί σε συλλογές εγγράφων συγκεκριμένης θεματολογίας.
- Πολλά **άχρηστα πρότυπα**, εισάγουν **θόρυβο**.
  - Αλλά το SVM φαίνεται να τον αντιμετωπίζει καλά.

Διαμόρφωση 4: χωρίς το WordNet

- Όπως η διαμόρφωση 3 αλλά **χωρίς** την ιδιότητα για τα **«καλύτερα» υπερώνυμα**.
- Έλεγχος αν οι επιδόσεις της διαμόρφωσης 3 **εξαρτώνται** από τη χρήση των υπερωνύμων του **WordNet**.
- Θα μπορούσαν να γίνουν **αντίστοιχοι έλεγχοι** και για τις **άλλες ιδιότητες** (*SN*, *WC*, *RK*) αλλά η εξάρτηση από το WordNet μας ενδιέφερε ιδιαίτερα, λόγω των Ελληνικών.

Πειραματικά αποτελέσματα

Με ερωτήσεις/κείμενα από τα TREC-2000, 2001.

μέθοδος	επιτυχία (%)	
Prager et al.	51.95 (80/154),	*60.15 (80/133)*
baseline (επανυλοπ.)	50.00 (80/160),	*58.39 (80/137)*
διαμ. 1 (προσομ. Joho)	61.88 (99/160),	*72.26 (99/137)*
διαμ. 2 (+ Wordnet)	63.13 (101/160),	*73.72 (101/137)*
διαμ. 3 (+ *n*-grams)	**72.50** (116/160),	***84.67*** *(116/137)*
διαμ. 4 (– Wordnet)	71.88 (115/160),	*83.94 (115/137)*

Διαμόρφ. 3 και 4: 200 αποκτηθέντα πρότυπα.

Εξαιρώντας ερωτήσεις χωρίς πρότυπα απαντήσεων.

Μεταβλητός αριθμός *n*-γραμμάτων

n-grams	διαμόρφ. 3 (%)	διαμόρφ. 4 (%)
100	68.13, *79.56*	70.00, *81.75*
200	**72.50, *84.67***	**71.88, *83.94***
300	68.75, *80.29*	71.25, *83.21*

- **Χωρίς ενδείξεις** ότι η απόκτηση περισσότερων προτύπων μπορεί να οδηγήσει σε βελτίωση.
- Χρειάζονται **περισσότερα πειράματα**, αλλά για αυτό απαιτείται **γρηγορότερη υλοποίηση** SVM.
  - Χρησιμοποιούμε την υλοποίηση του Weka (SMO του Platt) με τις προεπιλεγμένες παραμέτρους.

Συμπεράσματα 1ης ενότητας

- Νέα μέθοδος εντοπισμού μεμονωμένων αποσπασμάτων που περιέχουν απαντήσεις σε ερωτήσεις ορισμού.
  - **Συνδυάζει προηγούμενες** προσεγγίσεις ως ιδιότητες σε ένα SVM, με επιπλέον **αυτόματα παραγόμενα πρότυπα** *n*-γραμμάτων.
- Πειραματική αξιολόγηση 4 διαμορφώσεων με κείμενα και ερωτήσεις του TREC.
  - **Καλύτερα αποτελέσματα** από προηγούμενες μεθόδους.
  - Τα **παραγόμενα πρότυπα βοηθούν** σημαντικά.
  - Τα **υπερώνυμα** του WordNet συνεισφέρουν **ελάχιστα**.
    - Τουλάχιστον με τον τρόπο που την ενσωματώσαμε.

Διάρθρωση της ομιλίας

- Εισαγωγή:
  - Συστήματα ερωταποκρίσεων
  - Ερωτήσεις ορισμού.
- 1η ενότητα (πτυχιακή Μηλιαράκη, άρθρο COLING):
  - Προηγούμενες μέθοδοι για ερωτήσεις ορισμού.
  - Επιβλεπόμενη μέθοδος μάθησης.
  - Πειράματα με άρθρα εφημερίδων από το TREC.
- 2η ενότητα (πτυχιακή Γαλάνη, υπό δημοσίευση):
  - Πειράματα με ιστοσελίδες.
  - Μη επιβλεπόμενη μορφή της μεθόδου.

2η ενότητα: το ζητούμενο

- Είσοδος: ένας **όρος-στόχος** προς ορισμό και οι κορυφαίες **ιστοσελίδες** που επέστρεψε μια μηχανή αναζήτησης (Altavista) για τον όρο-στόχο.
  - Εξετάζουμε μόνο τις κορυφαίες 10 ιστοσελίδες, αντί για τα κορυφαία 50 έγγραφα της ενότητας 1.
  - Και σε αυτές μόνο τα πρώτα 5 παράθυρα του όρου-στόχου. Συντελεί σε μείωση της ανισορροπίας μεταξύ των κατηγοριών *ορισμός* και *μη ορισμός*.
- Έξοδος: μόνο **ένα απόσπασμα** των 250 χαρακτήρων.
  - Αντί των 5 αποσπασμάτων της ενότητας 1. Πιο δύσκολο.
  - Πιο εύκολη αξιολόγηση.
- Ορθή έξοδος αν περιέχει **αποδεκτό ορισμό**.
  - Τα αποσπάσματα κρίνονται από **ανθρώπους-κριτές**.
  - Δεν χρησιμοποιούμε πια κείμενα και απαντήσεις του TREC, οπότε δεν υπάρχουν πια πρότυπα απαντήσεων.

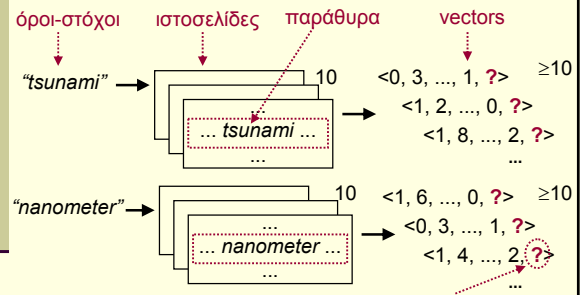
Παραδείγματα αποσπασμάτων

Για τον όρο «generic drug»:

✓ …time pharmacist questioning pharmacist consulting pharmacy drug compounding drug price what is generic drug what is a generic drug? a generic drug is one which is identified by its official chemical name rather than an advertised brand name…

✘ …there are new results for return to results glaxo wellcome's perspective on gatt patent debate; what is the generic drug industry's response to glaxo wellcome's? pr newswire; 12/5/1995 read the full article, get a free trial for…

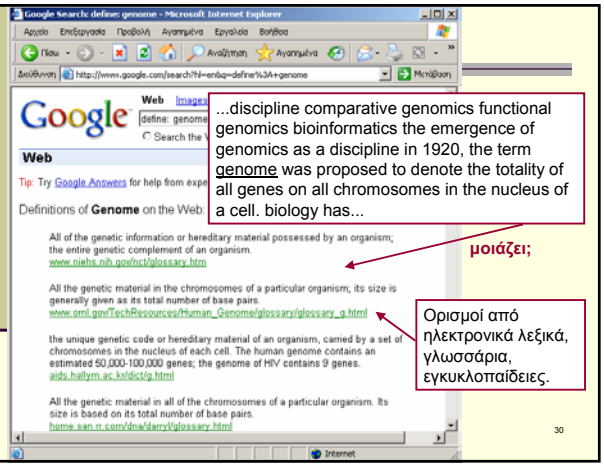
Το πρόβλημα κατά την εκπαίδευση



Πώς θα σημειώσουμε τις σωστές κατηγορίες (ορισμός, μη ορισμός); Δεν έχουμε πια στη διάθεσή μας πρότυπα απαντήσεων. Στα πειράματα της ενότητας 1, υπήρχαν 18.473 παράθυρα εκπαίδευσης!

Πιθανές λύσεις για την εκπαίδευση

- **Χειρωνακτική** κατάταξη χιλιάδων παραθύρων εκπαίδευσης.
- Εκπαίδευση σε ερωτήσεις και κείμενα **TREC**.
  - Όμως οι ιστοσελίδες διαφέρουν από τα άρθρα ειδήσεων του TREC.
- Εξεύρεση τρόπου **αυτόματης προσεγγιστικής κατάταξης** των παραθύρων εκπαίδευσης.
  - **Κατά την εκπαίδευση** μπορούμε να χρησιμοποιήσουμε όρους-στόχους από το ευρετήριο **ηλεκτρονικής εγκυκλοπαίδειας**.
  - Διαλέγουμε όρους-στόχους για τους οποίους έχουμε **πολλούς ορισμούς** από διαφορετικές εγκυκλοπαίδειες.
  - Κατατάσσουμε ένα παράθυρο εκπαίδευσης ως ορισμό όταν το λεξιλόγιό του **μοιάζει** αρκετά με εκείνο **πολλών αντίστοιχων ορισμών** εγκυκλοπαιδειών.



Πιθανές απορίες

- Γιατί να μη χρησιμοποιούμε **κατευθείαν το "define"** του Google;
  - Γιατί υπάρχουν πάντα όροι που **δεν περιλαμβάνονται** στα λεξικά, γλωσσάρια, εγκυκλοπαίδειες (π.χ., νέοι τεχνικοί όροι, ονόματα προσώπων, προϊόντων).
  - Η μέθοδός μας προσπαθεί να **συμπληρώσει** το "define" του Google βρίσκοντας ορισμούς σε κοινές ιστοσελίδες.
- Γιατί να μην **εκπαιδεύσουμε** τη μέθοδο κατευθείαν **στους ορισμούς** των λεξικών, γλωσσαρίων κλπ;
  - Γιατί οι εκφράσεις των λεξικών, γλωσσαρίων κλπ. **διαφέρουν** εν γένει από τις εκφράσεις των κοινών ιστοσελίδων.
  - Και οι ορισμοί των λεξικών κλπ. προσφέρουν μόνο παραδείγματα ορισμών (**θετικά**). Χρειαζόμαστε και παραδείγματα μη ορισμών (**αρνητικά**).

31

Πόσο μοιάζει;

ορισμοί από εγκυκλοπαίδειες

C

... tsunami ...

παράθυρο εκπαίδευσης

W

... tsunami ...

Για όλο το W: $sim(W,C) = \frac{1}{|W|} \cdot \sum_{i=1}^{|W|} sim(w_i, C)$

Για κάθε λέξη w_i του W: $sim(w_i, C) = fdef(w_i, C) \cdot idf(w_i)$

Ποσοστό ορισμών του C όπου εμφανίζεται η w_i.

Πόσο σπάνια είναι η λέξη στα Αγγλικά. Υπολογίζεται από τα έγγραφα του BNC.

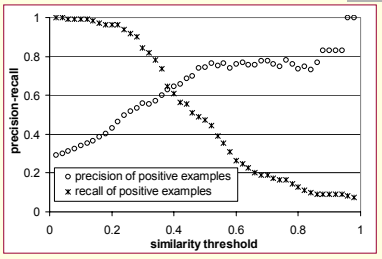
32

Ποιο κατώφλι ομοιότητας;

- Από ποια τιμή του ***sim(W,C)*** **και πάνω/κάτω** θα θεωρούμε το παράθυρο εκπαίδευσης **ορισμό/μη ορισμό**;
- Προκαταρκτικό πείραμα με **400 παράθυρα**.
  - Τυχαία επιλογή από τα παράθυρα που προέκυψαν από 130 ερωτήσεις TREC και τις ιστοσελίδες που επέστρεψε η Altavista.
  - Ένας άνθρωπος κατέταξε **χειρωνακτικά** σε ορισμούς και μη ορισμούς τα 400 παράθυρα.
  - Υπολογίσαμε το ***sim(W,C)*** για κάθε παράθυρο, χρησιμοποιώντας αντίστοιχους ορισμούς από το "define" του Google.

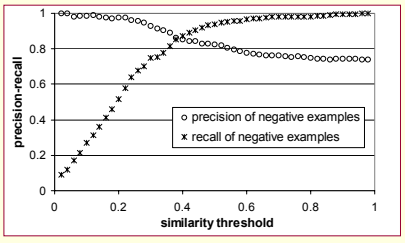
33

Θετική ακρίβεια και ανάκληση



- Θετική ακρίβεια (precision): Πόσα παράθυρα εκπαίδευσης που κατατάσσονται ως ορισμοί είναι όντως ορισμοί.
- Θετική ανάκληση (recall): Τι ποσοστό των παραθύρων εκπαίδευσης που είναι ορισμοί κατατάσσονται ως ορισμοί.

Αρνητική ακρίβεια και ανάκληση



- Αρνητική ακρίβεια: Πόσα από τα παράθυρα εκπαίδευσης που κατατάσσονται ως μη ορισμοί είναι όντως μη ορισμοί.
- Αρνητική ανάκληση: Ποσοστό παραθύρων εκπαίδευσης που είναι μη ορισμοί και κατατάσσονται ως μη ορισμοί.

35

Ποιο κατώφλι ομοιότητας; – συνέχεια

- **Δεν υπάρχει ένα κατώφλι** που να επιτυγχάνει ταυτόχρονα υψηλή θετική και αρνητική ακρίβεια.
- Χρησιμοποιούμε δύο κατώφλια: t_+ **και** t_-
  - Αν ***sim(W,C)*** > t_+ , κατατάσσουμε το παράθυρο εκπαίδευσης ως θετικό.
  - Αν ***sim(W,C)*** < t_- , κατατάσσουμε το παράθυρο εκπαίδευσης ως αρνητικό.

υψηλή αρνητική ακρίβεια: σχεδόν σίγουρα μη-ορισμός

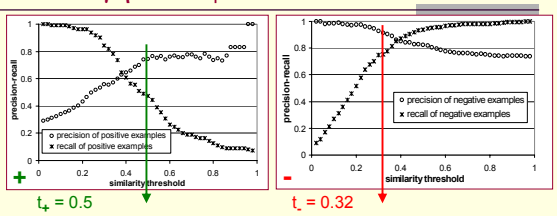
παράθυρα εκπαίδευσης που δεν χρησιμοποιούνται

υψηλή θετική ακρίβεια: σχεδόν σίγουρα ορισμός

sim(W,C) t_- t_+

36

Επιλογή κατωφλίων



$t_+ = 0.5$
θετική ακρίβεια: 0.72
θετική ανάκληση: 0.49

$t_- = 0.32$
αρνητική ακρίβεια: 0.92
αρνητική ανάκληση: 0.75

- Στα επόμενα πειράματα κρατάμε $t_+ = 0.5$ και επιλέγουμε το t_- ($0 \le t_- \le 0.34$), ώστε στα αυτόματα καταταγμένα παράθυρα εκπαίδευσης να διατηρείται η **αναλογία** ορισμών – μη ορισμών των 400 παραθύρων (0.37:1).
  - Για να αποφύγουμε μεροληψία (bias) υπέρ μιας κατηγορίας.

Πειραματική αξιολόγηση: συστήματα

- DEFQA-T: Το σύστημα της ενότητας 1 εκπαιδευμένο με **ερωτήσεις/κείμενα TREC**-2000, 2001:
  - 160 όροι-στόχοι, **3800 παράθυρα εκπαίδευσης**.
  - ≤10 κορυφαία έγγραφα, πρώτα ≤5 παράθυρα/έγγραφο.
  - 200 αποκτηθέντα πρότυπα.
- DEFQA-S: Ίδιο με το DEFQA-T αλλά εκπαιδευμένο σε **παράθυρα ιστοσελίδων** με *sim(W,C)*.
  - 480 όροι-στόχοι, **7200 παράθυρα εκπαίδευσης**.
  - Μπορούμε να παραγάγουμε αυτόματα όσα παράθυρα εκπαίδευσης θέλουμε!
  - Όροι-στόχοι από ευρετήριο www.encyclopedia.com.
- BASE-1: Πρώτο παράθυρο κορυφαίας ιστοσελίδας.
- BASE-R: Τυχαίο παράθυρο από 10·5 παράθυρα.

Πειραματικά αποτελέσματα

Με 81 νέους όρους-στόχους από www.encyclopedia.com.

μέθοδος	κριτής 1 (%)	κριτής 2 (%)	μέσος όρος (%)
BASE-R	14.81 (12/81)	14.81 (12/81)	14.81 (12/81)
BASE-1	14.81 (12/81)	12.35 (10/81)	13.58 (11/81)
DEFQA-T	25.93 (21/81)	25.93 (21/81)	25.93 (21/81)
DEFQA-S	**55.56 (45/81)**	**60.49 (49/81)**	**58.02 (47/81)**

- *K* = 0.86 (**ισχυρή συμφωνία** μεταξύ κριτών)
- Όλες οι διαφορές είναι **στατιστικά σημαντικές**.
  - Μονόπλευροι έλεγχοι διαφοράς αναλογιών (α = 0.001).
- Χειρότερα αποτελέσματα από την ενότητα 1 αλλά πιο δύσκολη εργασία (**μόνο 1 απόσπασμα/ερώτηση**).

Παρατηρήσεις

- Η DEFQA-S απάντησε σωστά περίπου **διπλάσιες** ερωτήσεις από την DEFQA-T.
  - **Παρά τον θόρυβο** στα δεδομένα εκπαίδευσης.
  - **Περισσότερα παράθυρα εκπαίδευσης**.
  - Ίσως τα πηγαίναμε καλύτερα με περισσότερα παράθυρα εκπαίδευσης.
- Πολύ **λιγότερα άσχετα** αποκτηθέντα πρότυπα.
  - Μάλλον θα τα πηγαίναμε καλύτερα με περισσότερα από 200 αποκτηθέντα πρότυπα.
- Πρότυπα **προσαρμοσμένα** σε εκφράσεις **ιστοσελίδων**.
  - π.χ. "*FAQ [στόχος]* ", "*home page [στόχος]* ", *"[στόχος] page*" , "*What is a [στόχος]* " , *"? A [στόχος]* "

Μελλοντικές κατευθύνσεις

- **Καλύτερα μέτρα ομοιότητας** (Γιακουμής, σε εξέλιξη).
  - Rouge από την αυτόματη παραγωγή περιλήψεων, συγκρίνει ακολουθίες λέξεων.
  - Συνδυασμός με μέθοδο κεντροειδούς (Cui et al.)
- **Περισσότερα παράθυρα εκπαίδευσης** και περισσότερα παραγόμενα **πρότυπα** (Γιακουμής).
- **Ομαδοποίηση** παρόμοιων αποσπασμάτων.
  - Ώστε να μην αξιολογούνται μεμονωμένα.
  - Διασαφήνιση εννοιών λέξεων ως υπο-προϊόν!
- **Διάταξη** (layout) ιστοσελίδων, γραμματοσειρές κλπ.
  - π.χ. προέρχεται το παράθυρο από ιστοσελίδα που μοιάζει με γλωσσάριο;
  - Ίσως χρησιμοποιείται στο "define" του Google.

Μελλοντικές κατευθύνσεις – II

- **Σημαντικότητα** (authority) των ιστοσελίδων.
  - Θα μπορούσε να είναι ιδιότητα του SVM.
- **Χρόνος ενημέρωσης/έκδοσης** ιστοσελίδων.
  - Π.χ. Ποιος είναι ο Κάρολος Παπούλιας;
- Ενσωμάτωση σε **συστήματα ερωταποκρίσεων**.
  - Κατάταξη ερωτήσεων σε κατηγορίες (Μαυροειδής).
  - Αναγνώριση/κατάταξη ονομάτων (Λουκαρέλλι).
  - Αυτόματη παράφραση (Παπαδημητρίου).
- Ενσωμάτωση σε **μηχανές αναζήτησης**.
- Εφαρμογή σε **αρχεία εφημερίδων** (Καρακατσιώτης).
  - Μέρος συστήματος ανάκτησης πληροφοριών για πρόσωπα.