ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΑΡΣΗ ΑΜΦΙΣΗΜΙΑΣ ΛΕΞΕΩΝ (ΑΠΟΣΑΦΗΝΙΣΗ ΕΝΝΟΙΑΣ ΛΕΞΕΩΝ) WORD SENSE DISAMBIGUATION

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Ιόνιο Πανεπιστήμιο**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons

Ασάφεια έννοιας λέξεων (Αμφισημία)

- 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)

Παραδοσιακή Προσέγγιση

- Εισαγωγή συντακτικών-σημασιολογικών περιορισμών στο πώς συνδυάζονται οι λέξεις
  - *Τρώω*: το υποκείμενο πρέπει να είναι ζωντανός οργανισμός και το αντικείμενο κάτι φαγώσιμο
  - *Πράσινος*: μπορεί να προσδιορίζει φυσικά αντικείμενα αλλά όχι αφηρημένες έννοιες
- Οι κανόνες αυτοί καλούνται επιλεκτικοί περιορισμοί (selectional restrictions)

Αποσαφήνιση Έννοιας Λέξεων (Word Sense Disambiguation)

- Η παραδοσιακή προσέγγιση μας επιτρέπει μόνο να ελέγξουμε αν κάτι είναι ή δεν είναι επιτρεπτό
- Στην πραγματικότητα δεν μπορούμε να ορίσουμε πλήρως τις επιτρεπτές σημασιολογικές ιδιότητες μιας έννοιας
  - *Έξυπνος άνθρωπος*
  - *Έξυπνη συσκευή*
  - *Έξυπνες κάρτες*
  - *Έξυπνα πλυντήρια*
- Οι στοχαστικές μέθοδοι μπορούν να βοηθήσουν προς αυτή τη κατεύθυνση βάσει ανάλυσης μεγάλων σωμάτων κειμένων (corpora)

Μοντέλο Μονογράμμου

- Η πιο απλή στοχαστική προσέγγιση είναι να μετρήσουμε πόσες φορές χρησιμοποιείται μία λέξη με την κάθε δυνατή έννοια 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

```
… <wrd sense=like2> like </wrd>    the
<wrd sense=bridge2> bridge </wrd>   of    …
```

N-gram model

- Χρησιμοποιώντας απλά unigrams θα διαλέγαμε ΠΑΝΤΑ την πιο συχνή έννοια για την κάθε λέξη

  (*γέφυρα* → κατασκευή)
- Επομένως πρέπει να λάβουμε υπόψη τα συμφραζόμενα
- Αν s_i είναι η έννοια (sense) της λέξης *i*
  - Bigrams: $P(s_n \mid s_{n-1})$
  - Trigrams: $P(s_n \mid s_{n-1}, s_{n-2})$

WSD - POS Tagging

- Υπάρχουν πολύ περισσότερες σημασιολογικές έννοιες από συντακτικές κατηγορίες
- 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 <u>οδοντίατρο</u> για να του φτιάξει τη <u>γέφυρα</u>.*
  - Η σημαντική λέξη βρίσκεται 6 λέξεις μακριά

Παράθυρο Συμφραζομένων

- 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 $P(w_5/s \mid w_1..w_9)$

(Naïve Bayes)

- Bayes rule: $P(A|B)=P(A)*P(B|A)/P(B)$
- Αν w είναι η κεντρική λέξη 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):
  $P(w_1..w_n \mid w/s) \rightarrow \Pi_{i=1..n}P(w_i \mid w/s)$
- Τελικό μοντέλο: $P(w/s \mid w_1..w_n)=P(w/s)* \Pi_{i=1..n}P(w_i \mid w/s)$

Εξαγωγή Μετρήσεων

- Δεδομένου ενός σχολιασμένου corpus και ενός μήκους παραθύρου λέξεων N εξάγουμε για την κάθε έννοια λέξης μετρήσεις:
- Ανεξάρτητες συμφραζομένων
  - P(*γέφυρα/κατασκευή*) = 0.3
  - P(*γέφυρα/πρ.οδοντικής*) = 0.1
- Εξαρτημένες από τα συμφραζόμενα
  - $P_N(w_i \mid w/s)$ = (# που το w_i υπάρχει μέσα στο παράθυρο όταν το *w/s* βρίσκεται στο κέντρο προς # που το *w/s* βρίσκεται στο κέντρο)
  - P_9(*οδοντίατρος* | *γέφυρα/πρ.οδοντικής*) = 0.2
  - Δηλ. η λέξη *οδοντίατρος* εμφανίζεται 2 φορές στις 10 σε ένα παράθυρο 9 λέξεων στο κέντρο του οποίου βρίσκεται η λέξη γέφυρα με την έννοια του προσθετικού οδοντικής

Παράδειγμα

- *P*(*γέφυρα/κατασκευή*) = 0.3
- *P*(*γέφυρα/πρ.οδοντικής*) = 0.1
- *P*9(*δόντι* | *γέφυρα/πρ.οδοντικής*) = 0.05
- *P*9(δόντι | *γέφυρα/κατασκευή*) = 0.001
- *P*9(*πυλώνας* | *γέφυρα/πρ.οδοντικής*) = 0.005
- *P*9(*πυλώνας* | *γέφυρα/κατασκευή*) = 0.01
- «Πήγα να βάλω μία γέφυρα στο μπροστινό μου δόντι.»
- *P1* = *P*(*γέφυρα/κατασκευή*)* *P*(*πήγα* | *γέφυρα/κατασκευή*)*..* *P*(*δόντι* | *γέφυρα/κατασκευή*)
- *P2* = *P*(*γέφυρα/πρ.οδοντικής*)* *P*(*πήγα* | *γέφυρα/πρ.οδοντικής*)*..* *P*(*δόντι* | *γέφυρα/πρ.οδοντικής*)
- Το πιο πιθανό είναι *P1* < *P2*

Λίστες Απόφασης (Decision lists)

- Εναλλακτικός τρόπος στοχαστικής αποσαφήνισης έννοιας λέξεων
- Με βάση ένα παράθυρο λέξεων εφαρμόζεται ένα σύνολο ελέγχων
- Ο πρώτος έλεγχος που επιτυγχάνει καθορίζει την έννοια της λέξης
- Αν αποτύχουν όλοι οι έλεγχοι επιλέγεται η πιο συχνή έννοια της λέξης

Παράδειγμα

- Αποσαφήνιση έννοιας «γέφυρα»

1. «κρεμαστή γέφυρα» → κατασκευή
2. «δόντι» μέσα στο παράθυρο → πρ.οδοντικής
3. «πυλώνας» μέσα στο παράθυρο → κατασκευή
4. «οδοντίατρος» μέσα στο παράθυρο → πρ.οδοντικής

Μάθηση Λιστών Απόφασης

- Καθόρισε τις πιθανότητες για όλες τις έννοιες λέξεων δεδομένων όλων των ελέγχων
- P(w/s | test)
  - # που ο έλεγχος ικανοποιείται και η λέξη έχει τη συγκεκριμένη έννοια προς # που ο έλεγχος ικανοποιείται με τη συγκεκριμένη λέξη
- Οι έλεγχοι ταξινομούνται σε φθίνουσα σειρά ανάλογα με την πιθανότητά τους

Μοντέλα

ΕΠΙΛΕΚΤΙΚΟΙ ΠΕΡΙΟΡΙΣΜΟΙ

- ☑ Αποδίδουν πολύ καλά όταν οι σημαντικές λέξεις συνδέονται μέσω συγκεκριμένων συντακτικών δομών
- ☑ Δεν επηρεάζονται από το πόσο απέχουν οι σημαντικές λέξεις μεταξύ τους
- ☑ Οι ιεραρχίες τύπων προσφέρουν πολλά πλεονεκτήματα
- ☒ Απαιτούν μεγάλη χειρονακτική εργασία
- ☒ Δυσκολεύονται στους ιδιωματισμούς
  - *Έφαγε τη σκόνη μου.*

ΣΤΟΧΑΣΤΙΚΑ ΜΟΝΤΕΛΑ

- ☑ Δεν επηρεάζονται από τις συντακτικές δομές
- ☑ Δεν απαιτούν μεγάλη χειρονακτική εργασία
- ☒ Το μήκος παραθύρου μπορεί να είναι πολύ περιοριστικό
- ☒ Δεν επωφελούνται από τις ιεραρχίες τύπων
- ☒ Κάποιες έννοιες λέξεων μπορεί να μην εμφανίζονται καθόλου ή ελάχιστα στο corpus εκπαίδευσης

Επιλεκτικούς Περιορισμούς

- Αποδίδουμε πιθανότητες σε έννοιες λέξεων που σχετίζονται με άλλες λέξεις μέσω συντακτικών δομών
  - Υποκείμενα, αντικείμενα ρημάτων
  - Επιθετικοί προσδιορισμοί
- Αντί να μετράμε πιθανότητες βάσει ενός παραθύρου Ν λέξεων, μετράμε πιθανότητες της μορφής:
  - *P*(K-PM=*βάζω* & ΑΝΤΙΚ=*γέφυρα/πρ.οδοντικής*)
  - *P*(K-PM=*βάζω* & ΑΝΤΙΚ=*γέφυρα/κατασκευή*)

Σημασιολογικοί Πόροι

- Λεξικά σε ηλεκτρονική μορφή (machine-readable dictionaries)
  - Longman dictionary of contemporary English (http://www.longman.com/ldoce/)
  - Collins English dictionary
- Θησαυροί (thesauri)
  - WordNet (www.cogsci.princeton.edu/wn/)

Λέξεις-Κλειδιά

- Λέξεις που όταν βρίσκονται στα συμφραζόμενα μιας λέξης αποσαφηνίζουν την έννοιά της
- Κάθε λέξη-κλειδί συνδέεται με μία έννοια
  - *δόντι, οδοντίατρος → γέφυρα/πρ.οδοντικής*
  - *πυλώνας, κρεμαστή → γέφυρα/κατασκευή*
- Δεν δίνουν πάντα λύση
  - *Ο οδοντίατρος πήγε στη γέφυρα του Ρίου.*
- Πρέπει κάποιος να ορίσει τις κατάλληλες λέξεις-κλειδιά για κάθε έννοια

Χρήση Λεξικών

- Μπορεί να γίνει χρήση των ορισμών των λεξικών για την εξαγωγή των λέξεων-κλειδιών
- Εναλλακτικά, μπορεί να γίνει σύγκριση των ορισμών διαφορετικών εννοιών δύο λέξεων
  - Επιλέγεται ο συνδυασμός με τη μεγαλύτερη επικάλυψη

Παράδειγμα: Χρήση Λεξικών

"pine cone"

pine 1. kinds of evergreen tree with needle-shaped leaves

2. waste away through sorrow or illness

cone 1. solid body which narrows to a point

2. fruit of certain evergreen trees

Μέγιστη επικάλυψη ορισμών

Λέξεις-κλειδιά: kinds evergreen tree needle-shaped leaves

Λέξεις-κλειδιά: waste sorrow illness

3η Προσέγγιση: Χρήση Θησαυρών

- Οι θησαυροί συνήθως κατηγοριοποιούν τις λέξεις σε θεματικές κατηγορίες
  - ιατρικός όρος, αθλητικά, κτλ.
- Οι θεματικές κατηγορίες ουσιαστικά είναι οι σημασιολογικές κατηγορίες
- Οι θεματικές κατηγορίες των συμφραζομένων προσδιορίζουν τη θεματική κατηγορία (έννοια) μιας λέξης

Βασικός Αλγόριθμος

- Για κάθε θεματική κατηγορία T_i
  - c_i ο αριθμός των λέξεων των συμφραζομένων που έχουν τη θεματική κατηγορία T_i ως μία από τις πιθανές
- Επέλεξε τη κατηγορία T_i με το υψηλότερο c_i
- Η έννοια είναι αυτή που αντιστοιχεί στο υψηλότερο T_i
- Πρόβλημα: Υπάρχουν περισσότερες έννοιες από θεματικές κατηγορίες

Corpus

- Η ύπαρξη σχολιασμένου corpus δεν είναι πάντα δυνατή
- Πολλοί άνθρωποι διαφωνούν για τις έννοιες συγκεκριμένων λέξεων σε συγκεκριμένα συμφραζόμενα
- Ο ορισμός ενός συνόλου εννοιών δεν είναι ποτέ πλήρης
  - Διαφορετικά επίπεδα εξειδίκευσης
- ΛΥΣΗ: Υπάρχουν μέθοδοι που δεν απαιτούν σχολιασμένο corpus!

Χωρίς Σχολιασμένο Corpus

- Αυτόματη ομαδοποίηση λέξεων με βάση στατιστικές μεθόδους (clustering)
- Οι ομάδες (clusters) των λέξεων μπορούν να θεωρηθούν ως «έννοιες»
- Συσσωρευτική ομαδοποίηση
  1. Δημιούργησε μία ομάδα για τη κάθε λέξη
  2. Συγχώνευσε τις δύο πιο όμοιες ομάδες
  3. Πήγαινε στο 2 μέχρι να ικανοποιούνται τα κριτήρια
- Κριτήρια:
  - Δημιουργία ενός καθορισμένου αριθμού ομάδων
  - Επίτευξη κάποιου ορίου που υποδεικνύει την ομοιότητα των ομάδων

Χαρακτηριστικά της Προσέγγισης

- Προβλήματα
  - ☒ Οι ομάδες δεν αντιστοιχούν σε πραγματικές έννοιες λέξεων
  - ☒ Ο αριθμός των ομάδων δεν είναι αντίστοιχος με τον αριθμό των εννοιών
- Αυτά τα προβλήματα μπορούν να μετατραπούν σε *πλεονεκτήματα*!
  - ☑ Οι ομάδες μπορούν να αντιστοιχούν σε εξειδικευμένες κατηγορίες εννοιών
  - ☑ Η μέθοδος μπορεί να χρησιμοποιηθεί για την *ανακάλυψη νέων εννοιών*!
    - Διαφορετικές έννοιες ως διαφορετικές χρήσεις της λέξης
    - Διαφορετικές εκφράσεις που ομαδοποιούνται με γνωστές έννοιες

Αξιολόγηση Συστημάτων WSD

- 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

Λίστες Εννοιών

- Τα αποτελέσματα αξιολόγησης εξαρτώνται από τη λίστα των διαφορετικών εννοιών
- Είναι δύσκολο να συγκρίνουμε δύο συστήματα που βασίζονται σε λίστες εννοιών με διαφορετικό βαθμό πολυπλοκότητας
  - Αν μια λέξη έχει δύο ισοπίθανες έννοιες, η αποσαφήνιση της σωστής έννοιας με πιθανότητα 90% είναι πολύ επιτυχής
  - Αν μια λέξη έχει δύο έννοιες με πιθανότητα 90% τη μία και 10% την άλλη, η αποσαφήνιση της σωστής έννοιας με πιθανότητα 90% είναι ασήμαντο επίτευγμα
- Λίστα βασικών εννοιών: 95% συμφωνία μεταξύ ανθρώπων
- Λίστα εξειδικευμένων εννοιών: 70% συμφωνία μεταξύ ανθρώπων

Έμμεση 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

Χρησιμότητα 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’