Σώματα Κειμένων και Εξόρυξη Ιατρικών Όρων: *C-value*

Κατερίνα Φραντζή

Τμήμα Μεσογειακών Σπουδών
Πανεπιστήμιο Αιγαίου

frantzi@rhodes.aegean.gr

συνεργασία επιστημονικών πεδίων

Πληροφορική

Γλωσσολογία

Ορολογία

Ιατρική / Βιολογία

οι Όροι

- οι όροι είναι η γλωσσολογική υλοποίηση ειδικών εννοιών

  protein kinase C, basal cell carcinoma
- η αναγνώριση των όρων δεν είναι απλή υπόθεση
- δεν υπάρχουν τυπικά κριτήρια για να ξεχωρίσουν οι όροι από τους μη όρους
- 'γνωστές' λέξεις μπορούν να είναι όροι

  WAS

  Wiskott-Aldrich Syndrome

  Ward Atmosphere Scale
- η συχνότητα εμφάνισης δεν είναι πάντα ο καλύτερος δείκτης

Αυξανόμενη ανάγκη για ορολογικούς πόρους

- Οι όροι είναι ο τρόπος της επιστημονικής επικοινωνίας, αναγκαίοι για την 'κατανόηση'
- Νέοι όροι (γονίδια, πρωτεΐνες, φάρμακα, χημικά, κτλ) δημιουργούνται συνεχώς
- Εκθετική αύξηση της βιοϊατρικής λογοτεχνίας
- Η επεξεργασία της ορολογίας προκαλεί 'μποτιλιάρισμα' στην επεξεργασία πληροφορίας στην βιολογία/ιατρική

Βήματα για την απόκτηση ορολογίας (term identification)

- Αναγνώριση όρων (term recognition)
- Κατηγοριοποίηση όρων (term classification)
- Ταίριασμα των όρων σε οντολογίες, βάσεις δεδομένων (term mapping)

Με την Αναγνώριση Όρων

- χειριζόμαστε άγνωστες λέξεις
- ενημερώνουμε λεξικά και οντολογίες
- Αποφεύγουμε την ορολογική σύγχυση

Αυτόματη Αναγνώριση Όρων (ATR)

- είναι το πρώτο βήμα για την απόκτηση ορολογίας
- βρίσκει τους όρους και τις ποικιλίες τους στο κείμενο
- ξεχωρίζει τους όρους από τους μη όρους

Εφαρμογές της ATR

- Ανάκτηση και Εξαγωγή Πληροφορίας – Information Retrieval and Extraction
- Κατηγοριοποίηση Κειμένων – Document categorisation
- Περίληψη - Summarisation
- …

ATR facts

1. Οι περισσότεροι όροι είναι πολυ-λεκτικοί
2. Οι φωλιασμένοι όροι είναι συχνοί
3. Τα ακρώνυμα είναι συχνά
4. Υπάρχει ποικιλία στους όρους
5. χρήση προθέσεων

η C/NC-value αντιμετωπίζει τα 1 και 2

Ποικιλία στους όρους (1/3)

Οι όροι εμφανίζονται με διάφορους τύπους

oestrogen	και	*estrogen*
amino acid	και	*amino-acid*
human cancer	και	*cancer in humans*
DNA	και	*deoxyribonucleic acid*

Κανονικοποίηση των ποικιλιών των όρων

retinoic acid receptor
- retinoic acid receptor
- retinoic acid receptors
- RAR, RARs

nuclear receptor
- nuclear receptor
- nuclear receptors
- NR, NRs

all trans retinoic acid
- all trans retinoic acid
- all-trans-retinoic acids
- ATRA, at-RA

Απλές ποικιλίες (2/3)

- ορθογραφικές
  - οριζόντιες, κάθετες παύλες (*amino acid* και *amino-acid*)
  - μικρά-κεφαλαία (*NF-KB* και *NF-kb*)
  - γραφή (*tumour* και *tumor*)
  - μεταγραφή σε άλλο αλφάβητο - transliterations (*oestrogen* και *estrogen*)
- μορφολογικές
  - κλίση (πληθυντικός, κτητικά)
- λεκτικές
  - συνώνυμα (*carcinoma* και *cancer*)

Απλές ποικιλίες

Αφορούν τα συστατικά των όρων και όχι τη δομή τους

Η κανονικοποίηση είναι άμεση

- Κανόνες για τη γραφή (παύλες, κτλ.)
- Κανόνες για τη μεταγραφή μεταξύ αλφάβητων (π.χ. *ph* → *f*; *oe* → *e*)
- Κανόνες για τα λήμματα (φαινόμενα κλίσης)
- Λεξικά συνωνύμων (λεκτικές ποικιλίες)

Πιο σύνθετες ποικιλίες (3/3)

- στη δομή
  - κτητική με τη χρήση προθέσεων (*clones of human* και *human clones*)
  - προθέσεις (*cell in blood* και *cell from blood*)
  - Συνδυασμός όρων - term coordination (*adrenal glands and gonads*)
- ακρώνυμα
  - (*nuclear factor kappa-B: NF-KB, NF-kb*, …)

Συνδυασμός Όρων

- Η δομή είναι ασαφής
- συνδυασμός (coordination) ή σύνδεση (conjunction) όρων;

παράδειγμα	*adrenal glands <u>and</u> gonads*
συνδυασμός	[*adrenal* [*glands <u>and</u> gonads*]]
σύνδεση όρων	[*adrenal glands*] *<u>and</u>* [*gonads*]

(N|A)⁺ CC (N|A)* N⁺

- chicken and mouse receptors
- cell differentiation and proliferation

Ακρώνυμα (1/3)

Ένας πολύ παραγωγικός τύπος ποικιλίας όρων

Ποικιλία στα ακρώνυμα (συνώνυμα)

- *nuclear factor kappa B*
- *NF-kappaB*
- *NF kappa B*
- *NF(kappa)B*
- *NFKB*
- *NF-KB,*
- *NF kB*

Ακρώνυμα (2/3)

Η ασάφεια στα ακρώνυμα (πολυσημία)
υπάρχει ακόμα και σε ελεγχόμενα λεξικά

GR για

glucocorticoid receptor
glutathione reductase

Ακρώνυμα (3/3)

ακρώνυμα	Ανεπτυγμένες φόρμες	κανονικοποιημένα
RAR alpha RAR-alpha RARA RARa	retinoic acid receptor alpha retinoic-acid receptor-alpha retinoic-acid receptor alpha	retinoic acid receptor alpha
APL	acute promyelocytic leukaemia acute promyelocytic leukemia	acute promyelocytic leukemia
9-c-RA 9cRA	9-cis-retinoic acid 9-cis retinoic acid	9 cis retinoic acid
RAR RARs	retinoic acid receptor retinoic acid receptors	retinoic acid receptor

Φωλιασμένοι Όροι
Nested terms (1/2)

Οι περισσότεροι όροι είναι πολύ-λεκτικοί
(περιορισμένο σύνολο συντακτικών δομών)

μέγιστοι και φωλιασμένοι όροι

αναγνώριση μερών του όρου
τα οποία επίσης αποτελούν όρους

enzyme inhibitors

angiotensin-converting enzyme inhibitors

Φωλιασμένοι Όροι (2/2)

Ορισμός σημασιολογικών σχέσεων μεταξύ των συστατικών (σαν πρώτο βήμα για το χτίσιμο οντολογιών)

Εσωτερική δομή όρου

[leukaemic [T [cell line]] Kit225]

Γλωσσολογία Σωμάτων Κειμένων

- Πραγματική χρήση της γλώσσας
- Μεγάλες συλλογές φυσικών κειμένων

Τα ΣΚ προσφέρουν

- συστηματική
- ακριβή
- πλήρη
- γρήγορη

επεξεργασία

Εφαρμογές των ΣΚ

- Μορφολογία
- Σύνταξη
- Σημασιολογία
- Φωνολογία
- Λεξικογραφία
- Διαλεκτολογία
- Ιστορικοσυγκριτική γλωσσολογία
- …

Εφαρμογές των ΣΚ

εκτός των καθαρά γλωσσολογικών

Εξόρυξη πληροφορίας:

- **Ορολογία**
- Δικαστική γλωσσολογία (forensic linguistics)
- Πολιτική γλωσσολογία (political linguistics)
- Παιδαγωγική
- Ψυχολογία
- …

C-value: μια υβριδική μέθοδος (1/3)

- η C/NC-value είναι μέθοδος για την εξαγωγή των πολύ-λεκτικών όρων
- Δίνει ιδιαίτερη προσοχή στους φωλιασμένους όρους (nested terms)
- υβριδική προσέγγιση εκμεταλλευόμενη
  - Γλωσσολογία (term formation patterns)
  - Στατιστική (ranking term candidates)
  - Πληροφορία από το «περιβάλλον» (contextual information, re-ranking)
- δίνει τιμές στους υποψήφιους όρους και τους ταξινομεί

C-value (2/3)

- συνολική συχνότητα της ακολουθίας λέξεων στο σώμα κειμένων
- συχνότητα της ακολουθίας λέξεων σαν μέρος μεγαλύτερων υποψήφιων όρων
- αριθμός αυτών των μεγαλύτερων υποψήφιων όρων
- μήκος του string (σε αριθμό λέξεων)

C-value (3/3)

adenoid cystic basal cell carcinoma
cystic basal cell carcinoma
ulcerated basal cell carcinoma
recurrent basal cell carcinoma
basal cell carcinoma


NaCTeM
The National Centre for Text Mining

National Centre for Text Mining (NaCTem)

http://www.nactem.ac.uk/

- Ανάκτηση Πληροφορίας

  Information Retrieval

- Εξαγωγή Πληροφορίας

  Information Extraction

- Διαχείριση όρων

  Term management

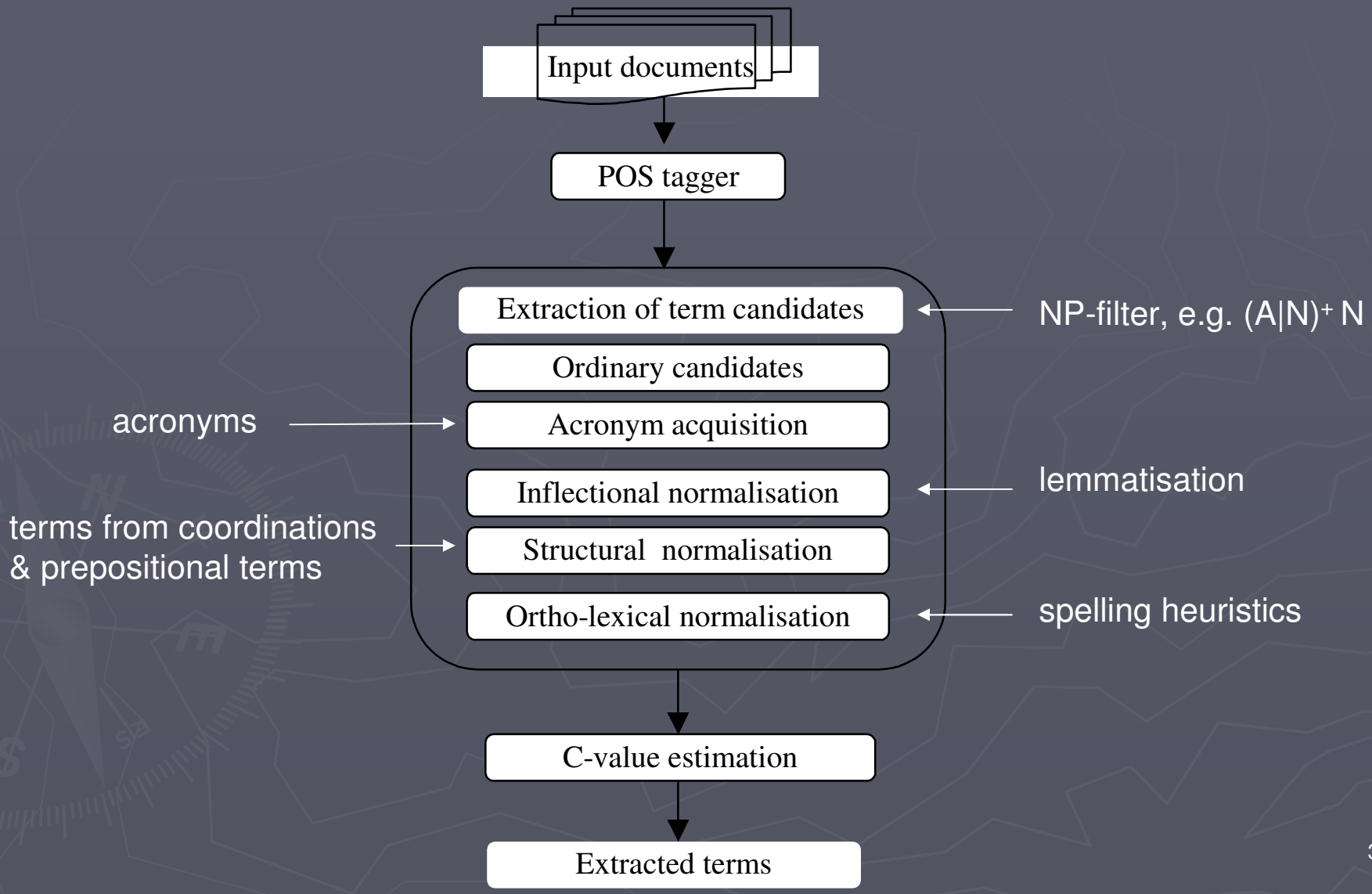
  - Biosciences
  - Humanities etc

TerMine

http://www.nactem.ac.uk/software/termine/

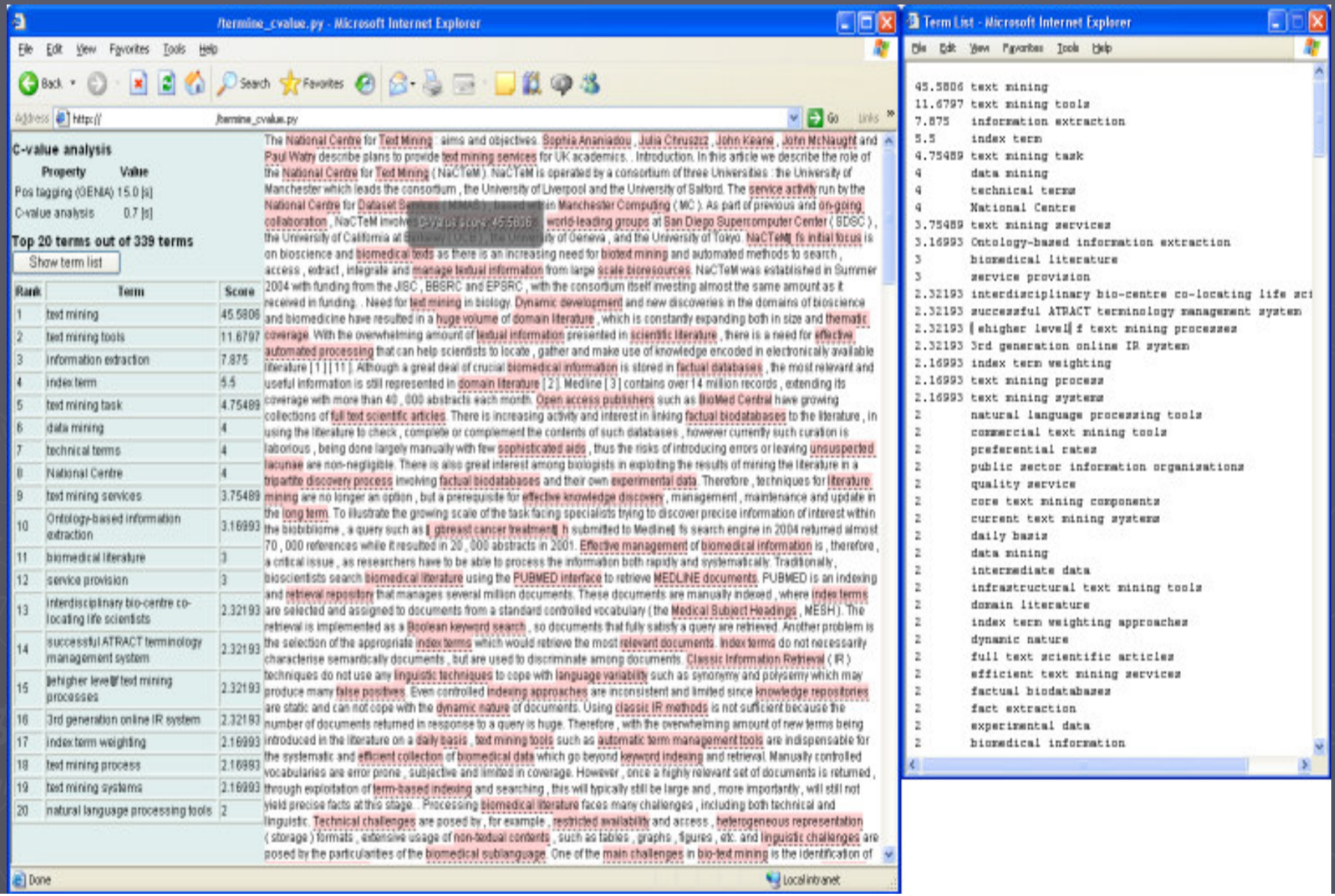
- Ένα σύστημα για τη διαχείριση της ορολογίας
- Χρησιμοποιεί την μέθοδο *C-value* για την εξαγωγή των όρων
- Σχετικά με την ταχύτητα της υλοποίησης:
  δοσμένου συνόλου από 1.3 εκατ. περιλήψεις
  από το MEDLINE (2GB κείμενο),
  το TerMine εξάγει 9.8 εκατ. υποψήφιους όρους καθώς και την τιμή τους (termhood)
  σε περίπου 10 λεπτά

ATR


Input documents
POS tagger
Extraction of term candidates
Ordinary candidates
Acronym acquisition
Inflectional normalisation
Structural normalisation
Ortho-lexical normalisation
C-value estimation
Extracted terms
NP-filter, e.g. (A|N)+ N
acronyms
lemmatisation
terms from coordinations & prepositional terms
spelling heuristics

TerMine (C-value) analysis demonstration

http://www-tsujii.is.s.u-tokyo.ac.jp/termine/



...

έχει γίνει έρευνα

η οποία έχει περάσει στην εφαρμογή

Επειδή όμως εμπλέκονται

Ορολογία

Πληροφορική

Ιατρική

Υπάρχει ακόμα αρκετός χώρος για πρόοδο