Εξαγωγή Περιλήψεων από Πολλαπλές Πηγές με Χρήση Οντολογιών

Γεώργιος Γιαννακόπουλος [1]
ggianna@iit.demokritos.gr

[1]Ινστιτούτο Τηλεπικοινωνιών και Πληροφορικής – Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού – Ε.Κ.Ε.Φ.Ε. Δημόκριτος σε συνεργασία με το Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων – Πανεπιστήμιο Αιγαίου

30 Νοεμβρίου 2007

Δομή της Παρουσίασης

Σκοπός της Παρουσίασης

- Στόχοι Διατριβής
- Παρουσίαση Αντικειμένου
- Επιτεύγματα
- Μελλοντική Δουλειά - Συζήτηση

Στόχοι Εκπόνησης Διδακτορικής Διατριβής

- Έρευνα και Υλοποίηση Μεθόδου Αξιολόγησης Περιλήψεων
- Πρόταση Τυποποίησης και Μέτρησης των Ποιοτήτων Περίληψης
- Περιγραφή και Υλοποίηση Συστήματος Εξαγωγής Περιλήψεων

Αξιολόγηση Συστημάτων Εξαγωγής Περιλήψεων

Η αξιολόγηση είναι δύσκολη διαδικασία. Υπάρχει και η έννοια της αξιολόγησης συστημάτων αξιολόγησης.

- Διακριτική ικανότητα
- Δύσκολα θέματα
- Άλλα

Αξιολόγηση Περίληψης

Τύποι αξιολόγησης:

- Εξωγενής

Αξιολόγηση Περίληψης

Τύποι αξιολόγησης:

- Εξωγενής
- Εγγενής

AutoSummENG

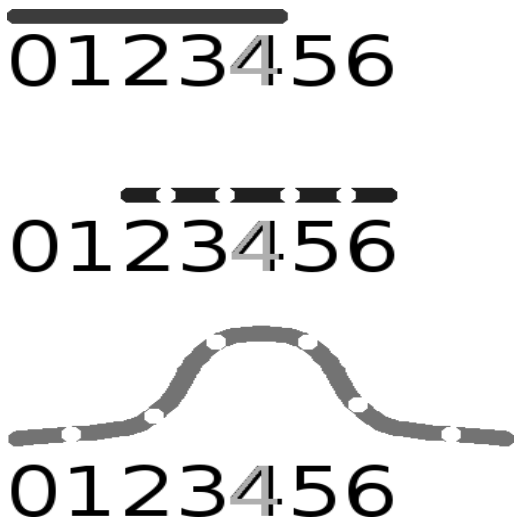
Αυτόματη Αξιολόγηση Συστημάτων Εξαγωγής Περιλήψεων με Χρήση N-Γραμμάτων

- Επιδόσεις εξίσου καλές με (ή και καλύτερες από) τα καλύτερα συστήματα του τομέα (ROUGE, ROUGE-BE, Pyramid).
- Καλύτερο από τα ROUGE-2, ROUGE-SU4 και με στατιστική υποστήριξη.
- Ελέγχθηκε στα δεδομένα τριών (3) ετών του DUC (2005, 2006, 2007).

AutoSummENG - Αναπαράσταση - Παράθυρα

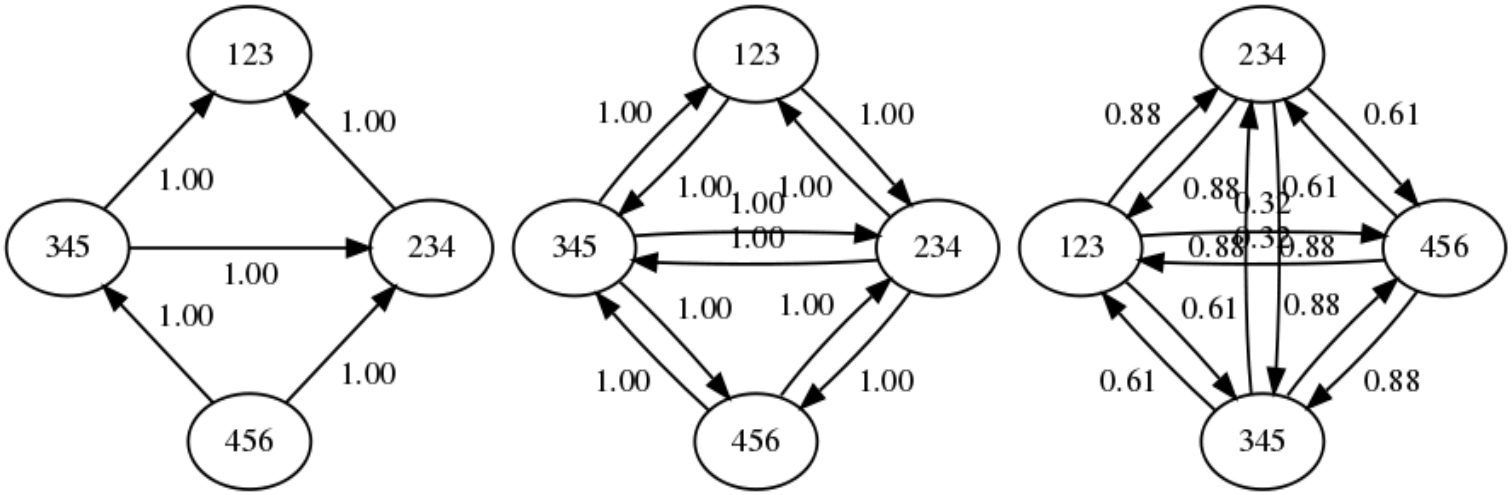
Γράφος ν-γραμμάτων λέξεων ή χαρακτήρων. Εξαγωγή βάσει παραθύρου.

Σχήμα: Τύποι παραθύρων ν-γραμμάτων (από πάνω προς τα κάτω): μη-συμμετρικό, συμμετρικό και gauss-κανονικοποιημένο συμμετρικό.



AutoSummENG - Αναπαράσταση - Τύποι Γράφων

Σχήμα: Γράφοι εξηγμένοι από τη συμβολοσειρά *123456* (από αριστερά προς τα δεξιά): μη-συμμετρικό, συμμετρικό και gauss-κανονικοποιημένο. N-γράμματα τάξης 3.



AutoSummENG - Αναπαράσταση - Συμπεράσματα 1

- Γράφος ν-γραμμάτων λέξεων ή χαρακτήρων.
- Καλύτερη επίδοση οι χαρακτήρες.
  Προβάλλεται η μέση επίδοση του συστήματος ανά αναπαράσταση στο DUC 2006.

Αναπαράσταση	*Μέση Συσχέτιση*
Γράφος ν-γραμμάτων λέξεων	0.478
Ιστόγραμμα ν-γραμμάτων λέξεων	0.767
Γράφος ν-γραμμάτων χαρακτήρων	0.915
Ιστόγραμμα ν-γραμμάτων χαρακτήρων	0.912

- Σύγκριση με χρήση δικής μας πρότασης για μετρική απόστασης γράφων ν-γραμμάτων.

AutoSummENG - Αναπαράσταση - Συμπεράσματα 2

- Χωρίς καμία γλωσσική πληροφορία. Απλή στατιστική ανάλυση.

AutoSummENG - Αναπαράσταση - Συμπεράσματα 3

- Χωρίς καμία γλωσσική πληροφορία. Απλή στατιστική ανάλυση.
- Πολλές εφαρμογές (π.χ. Computer-Assisted Stemmatology Challenge, φίλτρο ακατάλληλης αλληλογραφίας).

Αξιολόγηση - DUC 2006

- Ο στόχος: Ορθά εκφρασμένες περιλήψεις εις απάντηση ερωτημάτων εκφρασμένων σε ελεύθερη γλώσσα.

Αξιολόγηση - DUC 2006

- Ο στόχος: Ορθά εκφρασμένες περιλήψεις εις απάντηση ερωτημάτων εκφρασμένων σε ελεύθερη γλώσσα.
- Οι συμμετέχοντες: 35 αυτόματα συστήματα. 10 άνθρωποι.

Αξιολόγηση - DUC 2006

- Ο στόχος: Ορθά εκφρασμένες περιλήψεις εις απάντηση ερωτημάτων εκφρασμένων σε ελεύθερη γλώσσα.
- Οι συμμετέχοντες: 35 αυτόματα συστήματα. 10 άνθρωποι.
- Η αξιολόγηση: ROUGE-2, ROUGESU-4, BE σε σύγκριση με τέσσερεις περιλήψεις από ανθρώπους.

Αξιολόγηση - DUC 2006

- Ο στόχος: Ορθά εκφρασμένες περιλήψεις εις απάντηση ερωτημάτων εκφρασμένων σε ελεύθερη γλώσσα.
- Οι συμμετέχοντες: 35 αυτόματα συστήματα. 10 άνθρωποι.
- Η αξιολόγηση: ROUGE-2, ROUGESU-4, BE σε σύγκριση με τέσσερεις περιλήψεις από ανθρώπους.
- Η αξιολόγηση των μεθόδων αξιολόγησης: Όμοια κατάταξη με αυτή που ορίζουν οι ανθρώπινες βαθμολογίες (Συντελεστής συσχέτισης ταξινόμησης Spearman, Συντελεστής συσχέτισης Pearson, Συντελεστής συσχέτισης Kendall τ).

Πειραματικά αποτελέσματα 2006 - Συνολική Κατάταξη

Συντελεστής Spearman

Μετρική	*Spearman*	*Pearson*	*Kendall*
Rouge-2	0.767	0.836	
Rouge-SU4	0.790	0.850	
BE-HM	0.797	0.782	
AutoSummENG	*0.935*	*0.966*	0.804

Προσοχή: Πώς υπολογίζεται η επίδοση;

Πειραματικά αποτελέσματα 2005-2007

Συσχέτιση Spearman

Σύστημα	*2005*	*2006*	*2007*
AutoSummENG	**0.957347**	**0.9352756**	0.9348145
ROUGE-2	0.9571037	0.767	0.9291362
ROUGE-SU4	0.947454	0.790	0.9083835
ROUGE-BE	—	0.782	**0.9417847**

Εκ των Προτέρων Βελτιστοποίηση Παραμέτρων Αξιολόγησης

Παράμετροι:

- L_{min}, ελάχιστη τάξη ν-γράμματος.
- L_{MAX}, μέγιστη τάξη ν-γράμματος.
- D_{win}, εύρος παραθύρου.

Μέθοδος Βελτιστοποίησης Παραμέτρων

- Σύμβολα – Μη-σύμβολα
- Μοντέλο Θορυβώδους Καναλιού
- Εκτιμήτρια Επίδοσης
- Βελτιστοποίηση Παραμέτρων

Σύμβολα – Μη-σύμβολα

Σύμβολα Μεταφέρουν το νόημα του κειμένου.

Μη-σύμβολα Ακολουθίες χαρακτήρων που *έτυχε* να βρίσκονται διαδοχικά στο κείμενο.

Sq'hma: S'umbola

'permanent', 'permit', 'permits', 'persist', 'person', 'personal', 'personal computers'
'permi', 'permitt', 'pers', 'pers and', 'person kn'
'permit </HEADLINE>', 'permit program', 'permit approved'

Μοντέλο Θορυβώδους Καναλιού

- Σήμα $\equiv$ Σύμβολα
- Θόρυβος $\equiv$ Μη-Σύμβολα

Θέλουμε να ελαχιστοποιήσουμε την ποσότητα:

$$SN(L_{\min},\ L_{MAX}) = 10 \times \log_{10}\left(\frac{S(L_{\min},\ L_{MAX})}{N(L_{\min},\ L_{MAX})}\right) \qquad (1)$$

όπου $S(L_{\min},\ L_{MAX})$, $N(L_{\min},\ L_{MAX})$ είναι συναρτήσεις που επιστρέφουν μία εκτίμηση του σήματος και του θορύβου, για ένα εύρος τιμών των $L_{\min},\ L_{MAX}$.

Ποιότητα Εκτιμήτριας

Σχήμα: Συσχέτιση μεταξύ εκτίμησης και επίδοσης

JInsect Εργαλειοθήκη

JInsect Toolkit (www.ontosum.org)

- Πλαίσιο εργασίας για στατιστική ανάλυση κειμένων, βασισμένο *και* σε αναπαράσταση γράφων
- Περιέχει υλοποίηση του AutoSummENG και γραφικό περιβάλλον για την αξιολόγηση συστημάτων εξαγωγής περιλήψεων.
- Χρησιμοποιήθηκε για συμμετοχή στο Stemmatology Challenge (CASC). Πολύ καλά αποτελέσματα, που υποδεικνύουν γενική χρησιμότητα.

Τυποποίηση και Μέτρηση Ποιοτήτων Περιλήψεων

Ερωτήματα:

- Είναι απλή εφαρμογή των μετρικών που έχουν ήδη οριστεί με διαφορετικές παραμέτρους;

Τυποποίηση και Μέτρηση Ποιοτήτων Περιλήψεων

Ερωτήματα:

- Είναι απλή εφαρμογή των μετρικών που έχουν ήδη οριστεί με διαφορετικές παραμέτρους;
- Χρειάζεται η χρήση και μη στατιστικής γνώσης;

Στατιστική Εξαγωγή Θεματολογίας

Ιεραρχικό LDA

- Εξαγωγή στατιστικά οριζομένων θεμάτων, σε διαφορετικά επίπεδα ανάλυσης.
- Ουδετερότητα ως προς τη γλώσσα.

Στατιστική Εξαγωγή Θεματολογίας

Ιεραρχικό LDA

- Εξαγωγή στατιστικά οριζομένων θεμάτων, σε διαφορετικά επίπεδα ανάλυσης.
- Ουδετερότητα ως προς τη γλώσσα.
- Μη προφανής σημασιολογία.

Ορισμός Γραμματικής Κανονικότητας ή Ορθότητας

1. Δημιουργία γράφου για διάφορες τάξεις ν-γραμμάτων. Για κάθε τάξη:
   - Κάθε ν-γραμμα που συναντάται συνδέεται με τους χαρακτήρες που το ακολουθούν (κατανομή απόστασης).
2. Εκπαίδευση γράφου σώματος κειμένων (που ορίζουν την κανονικότητα).
3. Σύγκριση εκπαιδευμένου γράφου με το γράφο του κειμένου που εξετάζεται σε κάθε τάξη ν-γραμμάτων.

Εξαγωγή Γραμματικής Ορθότητας

- Σε επίπεδο ν-γραμμάτων χαρακτήρων
- Σε επίπεδο ν-γραμμάτων λέξεων
- Και στα δύο παραπάνω μαζί
- Σημαντικό εύρημα: Οι ανθρώπινες περιλήψεις δείχνουν να έχουν μικρότερη *γραμματική κανονικότητα* (με στατιστική σημασία) από τις αυτόματες. Τι σημαίνει αυτό;

Στατιστική Σύνθεση Περίληψης

- Εξαγωγή θεμάτων / συσχετίσεων
  - Ιεραρχικό LDA
  - Σημασιολογικό Ευρετήριο
- Εξαγωγή γραμματικής ορθότητας βάσει στατιστικού μοντέλου
  - Αλυσίδα Markov
  - Μοντέλα υψηλότερης τάξης
- Χρήση της από κοινού πιθανότητας για την παραγωγή νέου κειμένου, σχετικού με δεδομένη ερώτηση.

Εννοιολογικό Ευρετήριο

Χρήση

- Ανάλυση κειμένου εισόδου σε ν-γράμματα (με οποιαδήποτε μέθοδο)
- Τμηματοποίηση κειμένου με βάση στατιστικές μεθόδους
- Αναζήτηση τμήματος με χρήση του γράφου ν-γραμμάτων
- Εξαγωγή εννοιολογίας από τους επεστραμμένους κόμβους - Χρειάζεται σημασιολογική υποστήριξη (π.χ. *Οντολογία*)

Δημιουργία Συστήματος Εξαγωγής Περιλήψεων

Θα συνδυάζει τα συμπεράσματα από:

- Μοντελοποίηση Ποιοτήτων Περίληψης
- Μοντελοποίηση Πληροφοριακής Ανάγκης Χρήστη
- Χρήση Οντολογιών ως:
  - Υποστηρικτική γνώση για την εξαγωγή πληροφορίας
  - Πηγή παραφράσεων και συνωνύμων στην διαδικασία παραγωγής

Εξαγωγή Σημασιολογίας από Κείμενα

Εννοιολογικό Ευρετήριο - Δημιουργία

- Ανάλυση κορμού (corpus) κειμένων σε ν-γράμματα
- Δημιουργία γράφου μετάβασης ν-γραμμάτων
- Αντιστοίχιση κόμβων γράφου με έννοιες (οντολογία)
- Αντιστοίχιση ακμών γράφου με σχέσεις μεταξύ εννοιών (οντολογία)

Δημιουργία Περίληψης

- Μηχανική μάθηση για την εξαγωγή **συσχετίσεων** μεταξύ εννοιών
- Εφαρμογή της εξαγωγής συσχετίσεων σε συνάρτηση με ένα μοντέλο πληροφοριακής ανάγκης χρήστη, ως διαδικασίας εξαγωγής περίληψης
- Επιφανειακή αναπαράσταση της περίληψης **με χρήση κριτηρίων ποιότητας**

Στην Αξιολόγηση

- Μοντελοποίηση Διαδικασίας Εξαγωγής Περιλήψεων

Στην Αξιολόγηση

- Μοντελοποίηση Διαδικασίας Εξαγωγής Περιλήψεων
- Μελέτη Ποιοτήτων

Στην Αξιολόγηση

- Μοντελοποίηση Διαδικασίας Εξαγωγής Περιλήψεων
- Μελέτη Ποιοτήτων
- Ποσοτικοποίηση Ποιοτήτων

Στην Αξιολόγηση

- Μοντελοποίηση Διαδικασίας Εξαγωγής Περιλήψεων
- Μελέτη Ποιοτήτων
- Ποσοτικοποίηση Ποιοτήτων
- Αυτόματη Εξαγωγή Ποσοτήτων

Στην Παραγωγή

- Στατιστικές / ανεξάρτητες γλώσσας μέθοδοι σύνθεσης.
- Σύμφωνες με τις ανάγκες του χρήστη.

Εφαρμογή στη διαδικασία εξαγωγής περιλήψεων

1. Ορισμός Τομέα
2. Ανάλυση Θέματος και Ανάκτηση Πληροφορίας
3. Ανάλυση Δεδομένων
4. Παραγωγή και Αναπαράσταση Χαρακτηριστικών
5. Άθροιση και Συγχώνευση Πληροφορίας
6. Αναπαράσταση Περίληψης
7. Παραγωγή Περίληψης
8. Αναδιατύπωση Περίληψης
9. Αξιολόγηση Περίληψης

Ευχαριστώ

Γνωστική Προσέγγιση

- ὃνστιτυεντ Στρυστυρε
- Παραπηρασε υσινγ σψντασις συες
- Συπποσιτιον ερσυς Ασσερτιονς

Φιλολογική Προσέγγιση

- Οργάνωση Κειμένου
- Προθετικότητα

The Distribution of Symbols per Rank (Symbol Size) in the DUC 2006 corpus

Συνάρτηση Θορύβου

$$= \sum_{i=L_{\min}}^{M} |\text{Non-Symbols}_i| \tag{2}$$

Συνάρτηση Σήματος

$$P(s_r|s_{r-1}) = \begin{cases} \frac{1}{|\text{Symbols}_r|+|\text{Non-Symbols}_r|} \\ \frac{1}{|\text{Symbols}_{r-1}|+|\text{Non-Symbols}_{r-1}|} \times \frac{1}{|\text{Symbols}_r|+|\text{Non-Symbols}_r|} \end{cases} \quad \begin{matrix} \text{if } r = \\ \text{else.} \end{matrix}$$

Οπότε $w_r = 1/P(s_r|s_{r-1})$ (3)

'opou $|\text{Symbols}_r|$ είναι ο αριθμός συμβόλων τάξης r. Kai

$$W_r^0 = W_r \times \frac{|\text{Symbols}_r|}{\sum_{i=L_{\text{min}}}^{L_{\text{MAX}}} |\text{Symbols}_i|}$$