

Measuring Aboutness of an Entity in a Text

Marie-Francine Moens

Legal Informatics and Information Retrieval
Katholieke Universiteit Leuven, Belgium
marie-france.moens@law.kuleuven.be

Roxana Angheluta

Legal Informatics and Information Retrieval
Katholieke Universiteit Leuven, Belgium
anghelutar@yahoo.com

Patrick Jeuniaux

Department of Psychology
University of Memphis, USA
pjeuniaux@mail.psyc.memphis.edu

Rudradeb Mitra

Mission Critical, IT
Brussels, Belgium
rdm@missioncriticalit.com

Abstract

In many information retrieval and selection tasks it is valuable to score how much a text is about a certain entity and to compute how much the text discusses the entity with respect to a certain viewpoint. In this paper we are interested in giving an aboutness score to a text, when the input query is a person name and we want to measure the aboutness with respect to the biographical data of that person. We present a graph-based algorithm and compare its results with other approaches.

1 Introduction

In many information processing tasks one is interested in measuring how much a text or passage is about a certain entity. This is called *aboutness* or topical relevance (Beghtol 1986; Soergel 1994). Simple word counts of the entity term often give only a rough estimation of aboutness. The true frequency of the entity might be hidden by coreferents. Two entities are considered as coreferents when they both refer to the same entity in the situation described in the text (e.g., in the sentences: "Dan Quayle met his wife in college. The Indiana senator married her shortly after he finished his studies": "his", "Indiana senator" and "he" all corefer to "Dan Quayle"). If we want to score the aboutness of an entity with respect to a certain viewpoint, the aboutness is also obfuscated by the referents that refer to the chosen viewpoint and in which context the entity is mentioned. In the example "Dan Quayle ran for presidency", "presi-

gency" can be considered as a referent for "Dan Quayle". Because, coreferents and referents can be depicted in a graphical representation of the discourse content, it seems interesting to exploit this graph structure in order to compute aboutness. This approach is inspired by studies in cognitive science on text comprehension (van Dijk and Kintsch, 1983). When humans read a text, they make many inferences about and link information that is found in the text, a behavior that influences aboutness assessment. Automated aboutness computation has many applications such as text indexing, summarization, and text linking.

We focus on estimating the aboutness score of a text given an input query in the form of a person proper name. The score should reflect how much the text deals with biographical information about the person. We present an algorithm based on eigenvector analysis of the link matrix of the discourse graph built by the noun phrase coreferents and referents. We test the approach with a small set of documents, which we rank by decreasing aboutness of the input entity. We compare the results with results obtained by traditional approaches such as a normalized term frequency (possibly corrected by coreference resolution and augmented with other referent information). Although the results on a small test set do not pretend to give firm evidence on the validity of our approach, our contribution lies in the reflection of using graph based document representations of discourse content and exploiting this structure in content recognition.

2 Methods

Our approach involves the detection of entities and their noun phrase coreferents, the generation of terms that are correlated with biographical infor-

mation, the detection of references between entities, and the computation of the aboutness score. As linguistic resources we used the LT-POS tagger developed at the University of Edinburgh and the Charniak parser developed at Brown University.

2.1 Noun Phrase Coreference Resolution

Coreference resolution focuses on detecting “identity” relationships between noun phrases (i.e. not on is-a or whole/part links). It is natural to view coreferencing as a partitioning or clustering of the set of entities. The idea is to group coreferents into the same cluster, which is accomplished in two steps: 1) detection of the entities and extraction of their features set; 2) clustering of the entities. For the first subtask we use the same set of features as in Cardie and Wagstaff (1999). For the second step we used the progressive fuzzy clustering algorithm described in Angheluta et al. (2004).

2.2 Learning Biographical Terms

We learn a term’s biographical value as the correlation of the term with texts of biographical nature. There are different ways of learning associations present in corpora (e.g., use of the mutual information statistic, use of the chi-square statistic). We use the likelihood ratio for a binomial distribution (Dunning 1993), which tests the hypothesis whether the term occurs independently in texts of biographical nature given a large corpus of biographical and non-biographical texts. For considering a term as biography-related, we set a likelihood ratio threshold such that the hypothesis can be rejected with a certain significance level.

2.3 Reference Detection between Entities

We assume that the syntactic relationships between entities (proper or common nouns) in a text give us information on their semantic reference status. In our simple experiment, we consider reference relationships found within a single sentence, and more specifically we take into account relationships between two noun phrase entities. The analysis requires that the sentences are syntactically analyzed or parsed. The following syntactic relationships are detected in the parse tree of each sentence:

1) **Subject-object**: An object refers to the subject (e.g., in the sentence *He eats an apple*, *an apple* refers to *He*). This relationship type also covers

prepositional phrases that are the argument of a verb (e.g., in the sentence *He goes to Hollywood*, *Hollywood* refers to *He*). The relationship holds between the heads of the respective noun phrases in case other nouns modify them.

2) **NP-PP{NP}**: A noun phrase is modified by a prepositional noun phrase: the head of the prepositional noun phrase refers to the head of the dominant noun phrase (e.g., in the chunk *The nominee for presidency*, *presidency* refers to *The nominee*).

3) **NP-NP**: A noun phrase modifies another noun phrase: the head of the modifying noun phrase refers to the head of the dominant noun phrase (e.g., in the chunk *Dan Quayle’s sister*, *Dan Quayle* refers to *sister*, in the chunk *sugar factory*, *sugar* refers to *factory*).

When a sentence is composed of different subclauses and when one of the components of the first two relationships has the form of a subclause, the first noun phrase of the subclause is considered. When computing a reference relation with an entity term, we only consider biographical terms found as described in (2.2).

2.4 Computing the Aboutness Score

The aboutness of a document text D for the input entity E is computed as follows:

$$aboutness(D,E) = \frac{entity_score(E)}{\sum_{F \in \text{distinct entities of } D} entity_score(F)}$$

$entity_score$ is zero when E does not occur in D . Otherwise we compute the entity score as follows. We represent D as a graph, where nodes represent the entities as mentioned in the text and the weights of the connections represent the reference score (in our experiments set to 1 when the entities are coreferents, 0.5 when the entities are other referents). The values 1 and 0.5 were selected ad hoc. Future fine-tuning of the weights of the edges of the discourse graph based on discourse features could be explored (cf. Givón 2001). The edge values are stored in a link matrix A . The authority of an entity is computed by considering the values of the principal eigenvector of $A^T A$. (cf. Kleinberg 1998) (in the results below this approach is referred to as LM). In this way we compute the authority of each entity in a text.

We implemented four other entity scores: the term frequency (TF), the term frequency augmented with noun phrase coreference information (TFCOREF), the term frequency augmented with reference information (weighted by 0.5) (TFREF) and the term frequency augmented with coreference and reference information (TFCOREFREF). The purpose is not that the 4 scoring functions are mutually comparable, but that the ranking of the documents that is produced by each of them can be compared against an ideal ranking built by humans.

3 Experiments and Results

For learning person related words we used a training corpus consisting of biographical texts of persons obtained from the Web (from <http://www.biography.com>) and biographical and non-biographical texts from DUC-2002 and DUC-2003. For considering a term as biography-related, we set a likelihood ratio threshold such that the hypothesis of independence can be rejected with a significance level of less than 0.0025, assuring that the selected terms are really biography-related.

In order to evaluate the aboutness computation, we considered five input queries consisting of a proper person name phrase ("Dan Quayle" (D), "Hillary Clinton" (H), "Napoleon" (N), "Sadam Hussein" (S) and "Sharon Stone" (ST)) and downloaded for each of the queries 5 texts from the Web (each text contains minimally once an exact match with the input query). Two persons were asked to rank the texts according to relevancy, if they were searching biographical information on the input person (100% agreement was obtained). Two aspects are important in determining relevancy: a text should really and almost exclusively contain biographical information of the input person in order not to lose time with other information. For each query, at least one of the texts is a biographical text and one of the texts only marginally mentions the person in question. All texts except for the biography texts speak about other persons, and pronouns are abundantly used. The "Hillary Clinton" texts do not contain many other persons except for Hillary, in contrast with the "Dan Quayle", "Napoleon" and "Sadam Hussein" texts. The "Hillary Clinton" texts are in general quite relevant for this first lady. For "Napoleon" there is one biographical text on Napo-

leon's surgeon that mentions Napoleon only marginally. The "Dan Quayle" texts contain a lot of direct speech. For "Sharon Stone" 4 out of the 5 texts described a movie in which this actress played a role, thus being only marginally relevant for a demand of biographical data of the actress.

Then we ranked the texts based on the TF, TFCOREF, TFREF, TFCOREFREF and LM scores and computed the congruence of each ranking (R_x) with the manual ranking (R_m). We used the following measure of similarity of the rankings:

$$sim(R_x, R_m) = 1 - \frac{\sum_i |r_{x,i} - r_{m,i}|}{\text{floor} \frac{n^2}{2}} * 100$$

where n is the number of items in the 2 rankings and $r_{x,i}$ and $r_{m,i}$ denote the position of the i th item in R_x and R_m , respectively. Table 1 shows the results.

4 Discussion of the Results and Related Research

From our limited experiments we can draw the following findings. It is logical that erroneous coreference resolution worsens the results compared to the TF baseline. In one of the "Napoleon" texts, one mention of Napoleon and one mention of the name of his surgeon entail that a large number of pronouns in the text are wrongly resolved. They all refer to the surgeon, but the system considers them as referring to Napoleon, making that the ranking of this text is completely inversed compared to the ideal one. Adding other reference information gives some mixed results. The ranking based on the principal eigenvector computation of the link matrix of the text that represents reference relationships between entities provides a natural way of computing a ranking of the texts with regard to the person entity. This can be explained as follows. Decomposition into eigenvectors breaks down the original relationships into linear independent components. Sorting them according to their corresponding eigenvalues sorts the components from the most important information to the less important one. When keeping the principal eigenvector, we keep the most important information which best distinguishes it from other information while ignoring marginal information. In this way we hope to smooth some noise that is generated when building the links. On the other hand, when relationships that are wrongly detected

are dominant, they will be reinforced (as is the case in the “Napoleon” text). Although an aboutness score is normalized by the sum of a text’s entity scores, the effect of this normalization and the behavior of eigenvectors in case of texts of different length should be studied.

The work is inspired by link analysis algorithms such as HITS, which uses theories of spectral partitioning of a graph for detecting authoritative pages in a graph of hyperlinked pages (Kleinberg 1998). Analogically, Zha (2002) detects terms and sentences with a high salience in a text and uses these for summarization. The graph here is made of linked term and sentence nodes. Other work on text summarization computes centrality on graphs (Erkan and Radev 2004; Mihalcea and Tarau 2004). We use a linguistic motivation for linking terms in texts founded in reference relationships such as coreference and reference by biographical terms in certain syntactical constructs. Intuitively, an important entity is linked to many referents; the more important the referents are, the more important the entity is. Latent semantic indexing (LSI) is also used to detect main topics in a set of documents/sentences, it will not explicitly model the weights of the edges between entities.

Our implementation aims at measuring the aboutness of an entity from a biographical viewpoint. One can easily focus upon other viewpoints when determining the terms that enter into a reference relationship with the input entity (e.g., computing the aboutness of an input animal name with regard to its reproductive activities).

5 Conclusion

In this paper we considered the problem of ranking texts when the input query is in the form of a person proper name and when we are interested in biographical information. The ranking based on the computation of the principal eigenvector of the link matrix that represents coreferent and other referent relationships between noun phrase entities offers novel directions for future research.

6 Acknowledgements

The research was sponsored by the IWT-grant Nr. ADV/000135/KUL).

Table 1. Similarity of the system made rankings compared to the ideal ranking for the methods used with regard to the input queries.

	TF	TFCOREF	TFREF	TFCOREFREF	LM
D	0.33	0.00	0.33	0.00	0.50
H	0.33	0.50	0.33	0.33	0.66
N	0.66	0.33	0.66	0.66	0.33
S	0.83	0.66	0.66	0.66	1.00
ST	0.00	0.33	0.16	0.50	0.83

7 References

- Angheluta, R., Jeuniaux, P., Mitra, R. and Moens, M.-F. (2004). Clustering algorithms for noun phrase coreference resolution. In *Proceedings JADT - 2004. 7èmes Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-La-Neuve, Belgium.
- Beghtol, C. (1986). Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42(2): 84-113.
- Cardie C. and Wagstaff K. (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora*.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19: 61-74.
- Erkan, G. and Radev, D.R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22: 457-479.
- Givón, T. (2001). *Syntax. An Introduction*. Amsterdam: John Benjamins.
- Kleinberg, J.M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms* (pp. 668-677).
- Mihalcea, R. and Tarau, P. (2004). TextRank : Bringing order into texts. In *Proceedings of EMNLP* (pp. 404-411).
- Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science*, 45 (8): 589-599.
- Van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.
- Zha, H. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 113-120). New York : ACM.