Towards Focused Knowledge Extraction: Query-based extraction of structured Summaries

Besnik Fetahu
L3S Research Center
Hannover, Germany
fetahu@l3s.de

Bernardo Pereira Nunes
PUC-Rio
Rio de Janeiro, Brazil
bnunes@inf.puc-rio.br

Stefan Dietze
L3S Research Center
Hannover, Germany
dietze@l3s.de

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing–abstracting methods; I.2.7 [**Artificial Intelligence**]: Natural Language Processing–text analysis

General Terms

Algorithms, Experimentation, Theory

Keywords

POS patterns, text summarization, entity recognition

1. INTRODUCTION

A large part of Web resources consists of unstructured textual content. Processing and retrieving relevant content for a particular information need is challenging for both machines as well as humans. While information retrieval techniques provide methods for detecting suitable resources for a particular query, information extraction (IE) techniques enable the extraction of structured data [2, 3] and text summarization allows the detection of important sentences [1, 5]. However, neither IE nor summarization techniques do consider user interests when generating text summaries automatically.

This work presents an approach to extract focused knowledge, i.e. query-based and structured summaries according to particular user queries. In our case, summaries consist of structured data describing entities and their appearance contexts.

For identifying relevant statements and entities, we exploit a novel approach based on POS pattern analysis and entity recognition techniques.

2. PROBLEM DEFINITION

Briefly, we formalize the task of generating contextualized summaries and present examples for illustration. Let $D = \{d_1, d_2, \ldots, d_m\}$ be a set of documents and $T = \{t_1, t_2, \ldots, t_n\}$ a set of topics, where a topic is defined as a representation of most important terms from the corpus in D, formally defined as $t_i = \{w_1, w_2, \ldots, w_k\}$. We then define matrix $D \times T = [x_{ij}]_{(mn)}$, such that, $x_{ij} = o(d_i, t_j)$, for $i = 1 \ldots m \wedge j = 1 \ldots n$, where $o(d_i, t_j)$ is defined by a binary relation B indicating whether a document is related to a topic or not.

Now, let $Q = \{q_1, q_2, \ldots, q_z\}$ be a set of queries where $q_k = \{e_1, \ldots, e_v\}$ is a list of query terms. For instance, the user query "European+Union" results in the singleton term $e_1 =$ "European Union". The results is a subset of matching documents $D' \subset D$ and the set of topics $T' \subset T$, where $\forall t \in T', \exists d \in D' \wedge o(d, t) \in B$.

In what follows, we define the set σ as the union of POS tags from the terms in topic definitions from T' as $\rho = \cup_{(t \in T')} \omega(t)$ where $\omega \in \{NN, NNP, \ldots, VB, CD\}$ and the query terms from q_k as $\phi = \cup_{(e \in q_k)} e$, hence $\sigma = \rho \cup \phi$. Elements in σ are used to construct a square matrix which are added as row and column entries. The co-occurrence of two elements (σ_i, σ_j), for $i, j = 1 \ldots l$, computed for the documents in $D', P = [\delta(i, j)]_{lxl}$, e.g. $\sigma = \{NN, VB, \ldots,$ "European Union"$\}$.

Finally, a set of patterns $\Psi \in \{\psi_1, \ldots, \psi_y\}$ consists of a combination of elements from σ and a score assigned based on P. From documents in D' we define a set of sentences $S = \{s_{11}, \ldots s_{1v}, \ldots, s_{mv}\}$. As generated output from patterns in ψ and sentences in S, we define the focused summaries as $C = \{((s_{(i,j)}, \psi_k), (E, A))\}$ such that for $s_{(i,j)} \exists \psi_k \wedge f(s_{(i,j)}, \psi_k), f(s, \psi)$ is the match of sentence $s_{(i,j)}$ with pattern ψ. $E = \{e_1, \ldots, e_p\}$ and $A = \{a_1, \ldots, a_z\}$ are the set of entities and actions from sentence $s_{(i,j)}$ and $\forall e \in E, \exists e \in s$ and $\forall a \in A, \exists a \in s$.

3. OUR APPROACH

This work addresses extraction of *entities* and *actions* (a verb phrase that indicates an activity involving one or more entities) based on patterns that adapt to different user queries. The generation of patterns for a set of elements σ, consisting of query terms and their related entity terms found using query expansion and POS tags from terms in topic definitions in T', considers all their possible non-repetitive combinations.

Given a query q_k, we compute the conditional probabilities of the co-occurrences of the different entries (POS tag or query term) in the set of retrieved documents D' based on the generated matrix $P = [\delta(i, j)]_{lxl}$, see Section 2.

Thus, from P most probable patterns occurring in the set of retrieved documents for the elements in σ are computed. For each element a directed tree graph is modeled with all the possible combinations with other elements in σ (when there is a co-occurrence probability greater than zero in P).

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
Copyright 2013 ACM 978-1-4503-2038-2/13/05 ...$15.00.

The transition probabilities from one node to another represent the likelihood of terms from the document's text with a specific POS tag or query term appearing together.

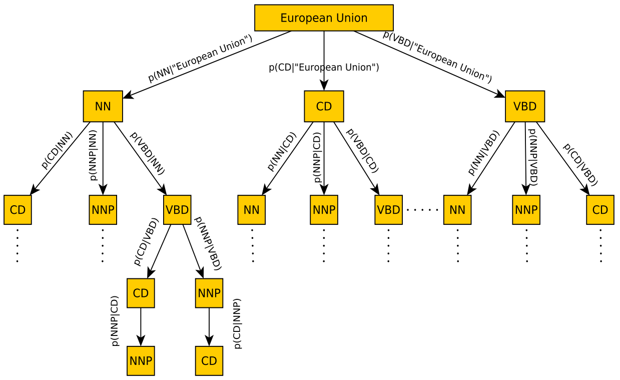

Figure 1: Pattern Generation approach using directed tree graphs.

The pattern scores are computed for all possible paths from the root node to the leaf nodes. Each path generates a pattern of variable number of elements, which is dependent on the co-occurrence of two elements from σ appearing in the retrieved documents.

The score of a pattern is computed as the marginal probability for the path that is present in a particular pattern. For instance, Figure 1 shows the pattern generation for the root node "European Union", computed as in Eq 1, where for the i-th row from matrix P are taken the probabilities and multiplied for each transition of different parent/child nodes afterwards.

$$\forall \psi \in \Psi, \psi_{score} = p(\sigma_i) \cdot \prod_{j=1}^{l} p(\sigma_{i,j}|\sigma_{i,j-1}) \tag{1}$$

As the number of generated patterns is large, we consider the top-10 patterns for each element with highest score computed as in Eq 1. Using the generated patterns individual sentences from the relevant documents are matched against one of the patterns.

A match is considered when a sentence contains an ordered set of terms having the same syntactical structure (POS tags and terms) as the patterns, we consider the relaxation of a full match and look for partial matches thus increasing coverage of the summaries.

4. EVALUATION AND RESULTS

To evaluate our approach we used ROUGE-n metrics [4]. ROUGE-n measures the coverage of the generated contextual summaries against human generated ones.

As a dataset we use the New York Times corpus, which contains approximately 40,000 news articles from 2007. Each article is annotated manually for entities occurring such as persons, locations and organizations, and contains a short abstractive summary.

We evaluate our approach in two directions: (i) *focused summary coverage*; and (ii) *focused summary appropriateness to a query*. For (i), we evaluate the query expansion on a set of queries taken based on their popularity and the expectedness of a broad coverage in our corpus. We also computed ROUGE-1, in terms precision/recall/F1, to evaluate the coverage of the generated focused summaries. As for (ii), we evaluated how well the generated summaries address specifically the query needs and how well they represent the query terms and the concepts implied by a query. For this, 17 participants assessed at least 20 summaries generated.

Based on the conducted evaluation, 76% of the generated summaries were relevant to the user queries and the concepts represented by them. This shows that our approach extracted focused knowledge with high precision by incorporating user queries for detecting the importance of specific POS tags. Furthermore, for ROUGE-1 we obtained high coverage precision over 25% and broad coverage with 32% in terms of recall. Our preliminary results are comparable to state of the art submissions in DUC[1]. Figure 2 shows the computed P/R/F1 for a small set of user queries.

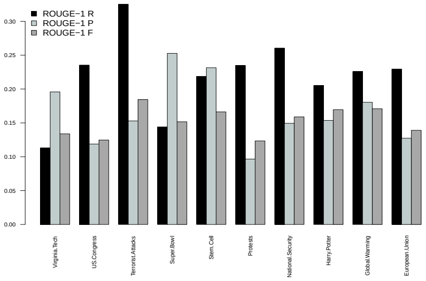

Figure 2: ROUGE-1 metric for different queries.

5. CONCLUSION

Our approach of focused knowledge extraction was applied for focused text summarization, demonstrating that it is able to produce targeted and structured summaries with high precision. Our main contributions are (a) the introduction of a novel POS tag pattern detection approach for relevance judgment of statements in unstructured texts, (b) the adaptation of a range of text and data processing techniques into a query-based document summarization approach and (c) the incremental population of a knowledge base describing entities and their appearance contexts.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, 2008.

[3] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545, 2011.

[4] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out Workshop*, pages 74–81, Barcelona, Spain, July 2004. ACL.

[5] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong. Integrating document clustering and multidocument summarization. *TKDD*, 5(3):14, 2011.

[1] http://duc.nist.gov/pubs.html