Summaries on the Fly: Query-Based Extraction of Structured Knowledge from Web Documents

Besnik Fetahu[1], Bernardo Pereira Nunes[1,2], and Stefan Dietze[1]

[1] L3S Research Center, Leibniz University Hannover, Germany
{fetahu,nunes,dietze}@L3S.de
[2] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil
bnunes@inf.puc-rio.br

Abstract. 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 and humans. While information retrieval techniques provide methods for detecting suitable resources for a particular query, information extraction techniques enable the extraction of structured data and text summarization allows the detection of important sentences. However, these techniques usually do not consider particular user interests and information needs. In this paper, we present a novel method to automatically generate structured summaries from user queries that uses POS patterns to identify relevant statements and entities in a certain context. Finally, we evaluate our work using the publicly available New York Times corpus, which shows the applicability of our method and the advantages over previous works.

Keywords: POS pattern analysis, knowledge extraction, text summarization, query-based summaries, entity recognition.

1 Introduction

The majority of Web resources consist of unstructured textual content. Due to the vast amount of information, filtering and adaptation of information to different user needs and contexts is crucial.

Information retrieval (IR) techniques facilitate the discovery and retrieval of relevant documents, often resulting in large sets of ranked documents shown to a user. When processing the retrieved documents, as part of such user queries, efficient methods are needed to enable users to quickly assess and judge the content of each document, in particular with respect to its relevance to the query.

Therefore, text summarization techniques aim at decomposing documents into its most important chunks like paragraphs, sentences, etc. Most prominent approaches on text summarization techniques rely on topic modeling methods [2], with each document belonging to one or more topics, and summarizing by detecting the importance of a sentence towards the defined topic. Despite the fact that text summarization approaches significantly reduce the amount of content,

they are not focused on the user interests. Hence, it often generates a generic summary of a textual document that might not reflect the user interests. Furthermore, after processing and detecting the most relevant concepts in a document, common text summarization techniques do not take advantage of the concepts found for representing the summaries in a structured form, which would improve reasoning over the structured text [1,3,28].

Information extraction (IE) approaches, specifically Named Entity Recognition (NER) tools and environments (e.g. GATE [7], DBpedia Spotlight[1], Alchemy[2], AIDA[3] or Apache Stanbol[4]), automatically generate structured data such as entities and their relationships [18] from unstructured Web resources, which would assist the information retrieval process.

In order to provide relevant information focused on particular user needs, we introduce a novel query-driven summarization and knowledge extraction approach based on POS pattern analysis, topic modeling and NER. Concisely, our approach exploits POS co-occurrence frequency from documents retrieved given a user query to summarize the results that match most frequent POS pattern. Additionally, we use DBpedia[5] and Freebase[6] as background knowledge to enrich, structure and disambiguate the concepts of each retrieved document.

As main contributions of this paper, we introduce a novel POS pattern detection approach for relevance judgment of statements in unstructured texts; adapt techniques of text and data processing into a *query-based document summarization* approach; create a new conceptual entity type based on the co-occurrence of certain POS tags, such as *noun phrases*; and, finally, the incremental population of a knowledge base for further reasoning. To the best of our knowledge, this is the first work that extracts focused and structured summaries, which satisfy given user queries and information needs. From now on, we refer to this approach as *focused knowledge extraction*.

The paper is structured as follows: Section 2 presents the related work on summarization and Section 3 introduces concepts used and formalizes the problem of focused knowledge extraction. Section 4 presents an overview and the preprocessing steps of our approach and Section 5 introduces the focused knowledge extraction for generating query-based summaries. Finally, in Section 6 we show the evaluation and the results of our work followed by a brief discussion and conclusions in Section 7.

2 Related Work

Most of the approaches for text summarization and extraction rely on combined methods. For instance, natural language processing (NLP) and information extraction (IE) techniques are usually used to generate extraction patterns [9],

[1] `http://spotlight.dbpedia.org`
[2] `http://www.alchemyapi.com`
[3] `http://adaptivedisclosure.org/aida/`
[4] `http://incubator.apache.org/stanbol`
[5] `http://dbpedia.org`
[6] `http://www.freebase.com`

while Latent Semantic Analysis (LSA) is combined with clustering techniques, such as Latent Dirichlet Allocation (LDA), to select representative textual content from texts [26].

As for IE approaches, the extraction of important pieces of information from textual contents is mainly based on entities and entity relations [9,17,10], where they use static patterns along with semantic and lexical features to achieve higher precision. The extraction of relations and events are usually performed in large sets of Web pages or data streams, such as Twitter[7] [21]. The approach on generating patterns for extracting relations is similar to ours with the difference that in our case instead of using fixed set of patterns, they are automatically generated based on the evidence provided by the retrieved documents for a specific user query.

Additional work on summarization [4,22,27,13] attempt on incorporating user query interests. However, they rely on naive heuristics of counting specific terms and defining manually extraction rules.

The field of Natural Language Processing (NLP) is a clear direction on leveraging the unstructured textual content, where the methods exploit the syntactic and semantic structure of languages used in resources. Related works on co-reference resolution depict the importance of an entity or part of sentence that can be implied for a specific context [15,20] and to resolve disambiguation of specific sentence parts. Similarly to our pattern generation approach, Hovy et al. [14] uses "Tree Kernels" to encode different needs of detecting events, relations and timestamps by incorporating POS tags, semantic types and other terms of interest. Moreover, SUMMONS [19] a summarization tool that builds templates for filling-in necessary information, and generates natural language as concise summary representation of the filled template. In our approach, we use co-reference to resolve ambiguities in the text.

A notable effort in text summarization tasks was performed by Blei et al. introducing the Latent Dirichlet Allocation (LDA) approach [2], which is based on a generative probabilistic model for topic construction. Particularly, we use LDA for generating clusters of a set of related topics. Apart from this, LDA is often used as a tool for summarization.

Other approaches on document clustering and summarization [26] rely on constructing document-term and sentence-term matrices using Latent Semantic Analysis (LSA). In this case, most important sentences selected based on generated eigenvalues from a non-negative matrix factorization are chosen as a base for language models. In this way, meaningful representations of clusters as sentences are generated rather than terms.

Following the same direction using LSA, Wan [25] considers subtopic creation from the main topic narrative text. Thus, sentences are measured for their relationship to the subtopics and presented as summaries for a particular subtopic. Similarly, Gong and Liu [12] consider IR and LSA techniques for ranking and identifying most important sentences as a means to construct summaries with broad coverage for a set of textual resources.

[7] `http://www.twitter.com`

Recent efforts from the semantic Web community consider the task of summarization from unstructured content [5,3,6,1], which are mainly based on the previously mentioned methodologies. Briefly, the approaches aim at summarizing the content into structured format such as Linked Data or as part of Ontology construction.

The method presented in this paper goes beyond the creation of text summaries and aims to generate structured context-based summaries. Although, previous semantic-based methods have partially addressed this issue, we incorporate specific user needs into an automatic pattern generation approach to extract only the information that fits the user query context.

3 Background

3.1 Concepts and Fundamentals

For the sake of clarity and to avoid confusion, we introduce concepts that are used throughout this work. An **action** is defined as a verb phrase that indicates an activity involving one or more **entities** as *subject/object*, whereas **entity** is a less restrictive concept compared to traditional NER approaches, and is not necessarily required to belong to one of the types (*people, location, organization,* etc.) or a newly defined **entity** *type iMisc* in Section 5.2.

Additionally, the previous concepts **action, entity** are also contextually defined. An **action context** captures additional information like *subject/object* as **entities** found in a specific context, whereas **entity context** contains additional descriptive information such as *adjectives, quantities, etc.*

3.2 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 result is a subset of matching documents $D' \subset D$ and the set of topics $T' \subset T$, where $\forall t \in T', \exists d \in D' \wedge o(d, t) \in B$. Note that, we also perform a query expansion step for each $q_k \in Q$, however, to preserve the clarity of the definition, we assume that the new terms introduced by the query expansion method are already considered in Q.

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. $\sigma = \{NN, VB, \ldots, \text{“European Union”}\}$.

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 $\forall a \in A, \exists a \in s$.

4 Overview and Running Example

This section presents the overall workflow of our focused knowledge extraction approach based on a running example. Fig. 1 shows the whole process starting from the user query input. Indeed, the user plays a central role in the generation of the summary, since the resulting summary is based on the user query terms.

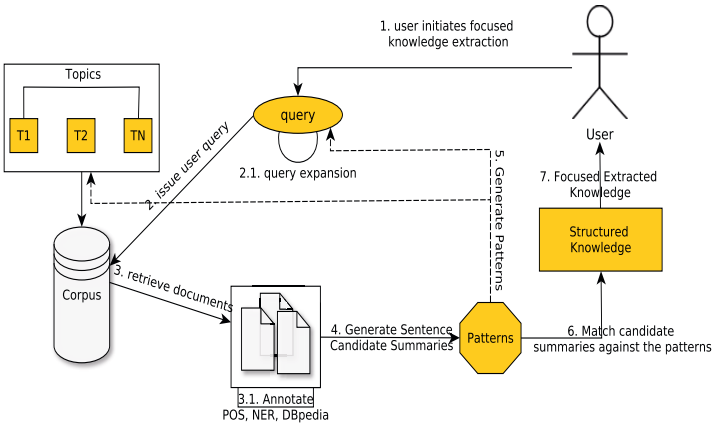

Fig. 1. Focused Knowledge Extraction Workflow

Let $q_1 = \{\text{“European Union”}\}$ be a query term where $q_1 \in Q$ issued by the user. Thus, the query term q_1 is processed and expanded using reference datasets, which results in new terms $q_k \in Q$. For instance, the query expansion for “European Union” results in $Q = \{\text{“European Union”}, \text{“European Union member economies”}, \text{“G20 nations”}, \ldots, \text{“International Organizations of Europe”}\}$.

The query expansion is performed for each query term provided by the user, where based on reference datasets, such as DBpedia and Freebase, related terms are automatically added to the list of user query terms Q. The terms added to

the list Q are labels (`rdfs:label`) from the directly related **entities** in such reference datasets, explained in detail in Section 5.1. The query expansion aims at improving recall and might be useful to disambiguate a particular user query. The disambiguation occurs when the query has multiple terms, which facilitates the identification of the user context.

Once the query terms are expanded, a set of relevant documents is retrieved, according to these terms. Since the corpus is pre-processed, annotated with POS tags and co-reference resolution applied, the task is synthesized to generate a set of patterns scored for their likelihood of appearance on the set of the retrieved documents.

Thus, in the case of the query Q, the top patterns generated is [$JJ \rightarrow VB \rightarrow$ "European Union" $\rightarrow RB$]. The set of topics is defined by the 1000 most representative topic terms extracted from the corpus. The set of the topics are selected using a topic modeling tool based on LDA [2] and annotated with POS tags.

As a result, we obtain all the documents and topics that serve as input to generate the summaries focused on the extracted knowledge and based on user queries. The example below shows a generated summary for query "European Union", in blue color are shown the **entities**, while in red the **actions**.

Bulgaria → joined → European Union, on Monday → helping to end → geographic divisions → left → cold war → extending → borders of the now 27-member bloc eastward to the Black Sea.

5 Focused Knowledge Extraction: Query-Based Summaries

In this section, we describe in details our approach of generating structured and focused summaries for specific user queries. For the focused summaries we propose an entity-based view which emphasizes **entities** and the contexts and **actions** in which they appear. In the following subsections are explained in details the necessary steps towards extracting and generating the focused summaries.

5.1 Query Expansion and Co-reference Resolution

The process of *query expansion* analyzes separately each query term for matching entities in the reference dataset DBpedia, and expands with related entities that are directly connected from all properties and assigned to the original query term. Moreover, the related query terms are extracted from the related entities using their label (`rdfs:label`). For instance, the query term "European Union" is considered as a singleton term if it is indicated as a cunjunction of terms. Finally, the query is reformulated as the disjunction of the original terms and the ones found during *query expansion*. However, this step can be exploited in addition also as a *query refinement* process by considering the conjunction rather than the disjunction.

Whereas co-reference resolution aims at resolving ambiguities of terms e.g "the president of the European Union" can be resolved to "Herman Van Rompuy", using Stanford's NLP tool [15,20].

5.2 *iMisc* Entity Type Definition

Determining the entity *type* is important for our approach, thus for named entity recognition we rely on the approach in [11], which detect annotation types such as *person, location, organization, date*. However, in many cases detecting the entity type is not possible, hence we rely on a *term matrix* which computes co-occurrence term frequencies of noun phrases among a set of previously analyzed and annotated documents based on the approach in [24,23], and recognizes named **entities** of *type iMisc* to distinguish from the other *types*.

An entity of type *iMisc* consists of terms which co-occur and can be formalized as the following: $entity[iMisc] = \bigcup_{i=1}^{k} \text{co-occur}(term_i, term_{i+1})$, where, in our case the maximum value for k was found to be 3 (indicating 3 terms that co-occur).

5.3 Automated Pattern Generation

One of the main challenges on creating user-query based summaries, is the extraction of **entities** and **actions** relying on patterns that adapt automatically to the intent of a user and set of retrieved documents. A pattern consists of a combination of items from the set σ that co-occur in a set of retrieved documents (see Section 3.2), with POS tags extracted from the annotation of topic definition terms and query terms.

Note that the set of POS tags is limited only to the topic definition terms (as representative for the set of retrieved documents), and ignore other POS tags not related to the topic definition terms. Thus, for a set of pairs of POS tags and query terms (σ), all non-repetitive combinations are considered to construct patterns for a given user query.

The combinations are represented in a symmetric matrix $P = [\delta_{(i,j)}]_{l \times l}$ in Eq. 1, hence, as rows and columns items from the set σ. The matrix is computed for each issued query and each entry $(\delta_{(i,j)})$ of the matrix represents the conditional probabilities of two items from σ co-occurring in the set of retrieved documents D'.

For instance, consider again our running example with the query "European Union" (referred with the acronym **EU**), which after the *query expansion* step results in the set of query terms $Q = \{$"European Union member economies", "G20 nations", ..., "International Organizations of Europe"$\}$. The resulting matrix is as follows:

$$P = \begin{bmatrix} & CD & VBD & \dots & NN & EU \\ CD & p(CD|CD) & p(VBD|CD) & \dots & p(NN|CD) & p(EU|CD) \\ VBD & p(CD|VBD) & p(VBD|VBD) & \dots & p(NN|VBD) & p(EU|VBD) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ NN & p(CD|NN) & p(VBD|NN) & \dots & p(NN|NN) & p(EU|NN) \\ EU & p(CD|EU) & p(VBD|EU) & \dots & p(NN|EU) & p(EU|EU) \end{bmatrix} \quad (1)$$

Given the resulting matrix P in Eq 1, we compute all possible combinations of patterns, supported by evidence from the set of retrieved documents. The problem of automatically generating patterns is modeled as a *directed tree graph*, thus, a pattern represents a *path* from a **root node** to a **leaf node**.

From each element in σ for a query a *directed tree graph* is modeled with all possible combinations with other element in σ (when the conditional probability between the two elements is than zero in P). The transition probabilities from one node to another represent the likelihood of those elements from the document's text with a specific POS tag or query term appearing together. Therefore, each path from the **root node** to one of the **leaf nodes** represents a pattern of variable number of elements.

The pattern scores are computed for the path from the **root node** to one of the **leaf nodes**. The score of a pattern represents the marginal probability of the probability of two nodes in the path co-occurring in the retrieved documents.

In more details, for an σ_i considered as the root node ("European Union") of the *directed tree graph*, as shown in our example in Fig. 2. The score of the pattern having as a root node "European Union" is computed as in Eq. 2, where for the i-th row in matrix P probabilities for each parent/child node transition are multiplied. Finally, the higher the score of the pattern the more important the pattern is, conveying important information about the most representative syntactical and semantical structures of a document.

$$\forall \psi \in \Psi, \psi_{score} = p(\sigma_i) \cdot \prod_{j=1}^{l} p(\delta_{i,j} | \delta_{i,j-1}) \tag{2}$$

To reduce the large number of detected patterns, we retain only the top-10 high scoring patterns as computed in Eq. 2.

Table 1 shows a small subset of patterns with highest scores generated for our running example. Using the generated patterns, individual sentences from the retrieved documents are *matched* against one of the patterns, and are further considered for generating focused summaries. A *match* is considered when a sentence contains an ordered set of terms having the same syntactical structure (ignoring POS tags that are not found in the topic definition terms) as a pattern, we consider the relaxation of a full match and look for partial matches thus increasing coverage of the summaries.

5.4 Contextual Structure of Extracted Knowledge

A necessary and important step after finding sentences decomposed from the retrieved documents is extraction of the knowledge as a pre-condition for generating focused summaries. As indicated in Section 5, our summaries provide an entity centric view, following the RDF schema visualized in Fig. 3.

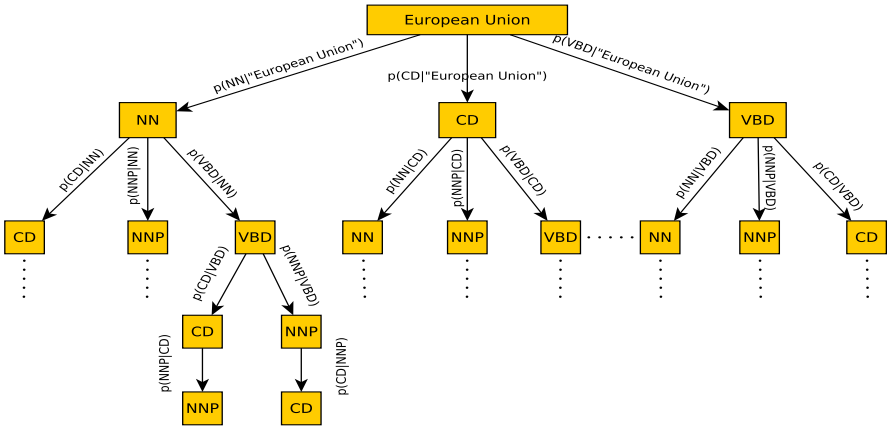

Fig. 2. Pattern Generation approach using directed tree graphs

Table 1. Automatically generated extraction patterns

Generated Patterns	Pattern Score ψ_{score}
$JJ \rightarrow VB \rightarrow$ "European Union" $\rightarrow RB$	$5.71E-09$
$JJ \rightarrow NN \rightarrow RB \rightarrow VB \rightarrow EU$	$4.63E-09$
$VB \rightarrow$ "European Union" $\rightarrow JJ \rightarrow RB$	$2.86E-09$
$VB \rightarrow$ "European Union" $\rightarrow JJ \rightarrow NN \rightarrow RB$	$1.16E-09$
"European Union" $\rightarrow JJ \rightarrow NN \rightarrow RB \rightarrow VB$	$6.99E-10$

In Fig. 3, similar as in [8] we consider several structures describing concepts introduced in Section 3.1. We separate the defined structures into two categories ***global*** and ***local***, explained in more details below.

Global structures such as **entity** and **action** capture relevant information about these concepts, disregarding their context. Only the description and the *type* of an **entity** as defined using standard NER tools[8] and the defind *iMisc type*. While, for an **action** the *state* as the verb tense is extracted and used as an indicator of whether the *action* is completed or an ongoing/future activity. Additionally, **entities** are enriched using DBpedia Spotlight with reference datasets like DBpedia, and a link (`owl:sameAs`) is provided to the reference instance in DBpedia.

Local structures like **entity-context** and **action-context** capture *contextual* information about the two global structures **entity** and **action**. With respect to **entity-context**, attributes (terms of POS tag *adjective*) and features like quantifiers (terms of POS tag as *cardinal number*) are captured for an **entity** describing for a specific context. Whereas for **action-context** we consider

[8] `http://nlp.stanford.edu/software/corenlp.shtml`

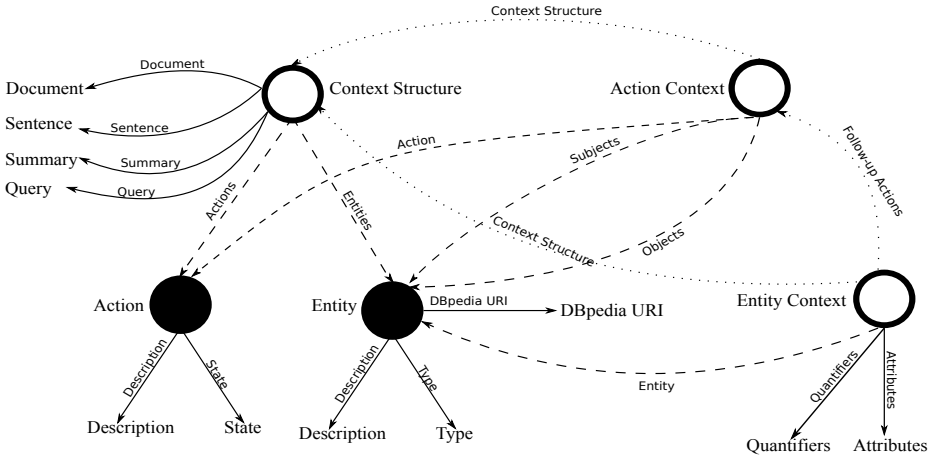

Fig. 3. Focused Knowledge Extraction RDF Schema

subject/object (**entities** belonging to the same context) as context specific information with which an **action** is linked. Finally, the *context structure* captures information about the source of information, matching pattern, along with the source document and user query.

The proposed RDF schema for representing and storing the focused summaries offers the functionality of viewing **entities** appearing in different context, showing the perspectives and their involvement for different queries. While additional information obtained after **entity** enrichment provides an interlinking mechanism to other data sources that lead to inferring of new knowledge for focused summaries.

Final aim of our *focused knowledge extraction* is constructing a publicly available knowledge base of summaries generated for different corpora and contexts over time, which will be incrementally populated and enriched. Access to the schema, the RDF dataset and other related tools and evaluation is available from a dedicated Website[9].

6 Evaluation and Results

In this section, we present a thorough evaluation of our approach followed by results and discussion. Concisely, the automatically generated summaries by our method are compared against abstractive manually created summaries. An "abstractive summary" is the summary that does not necessarily contain a similar syntactical structure as the original document, but covers its main concepts. The relevance of the automatically generated summaries for a given query against the original manually created summaries for each document in our corpus was assessed by humans and also using ROUGE [16].

[9] `htttp://l3s.de/~fetahu/QueryBased_Summaries/`

6.1 Dataset

As for the dataset, we used a subset of the New York Times (NYT) corpus, which contains 40,000 articles and its manually generated abstractive summaries from 2007. The articles are manually annotated with **entities** such as persons, locations and organizations[10]. In general, the length of the summaries from the NYT corpus ranges from 1 to 3 sentences. These summaries are used as gold-standard to measure the coverage of the automatically and contextualized summaries generated by our approach.

6.2 Evaluation Process

The evaluation is divided into two steps: (1) focused-summary appropriateness to user queries; and (2) focused-summary coverage.

The evaluation of step (1) aims at measuring how well an automatically generated summary represents the query terms and concepts implied by the query. In this evaluation, we created a questionnaire where we showed to the participants the query terms used to retrieve the documents and the automatically generated summaries. The participant has also access to the original summary and the document content. For this evaluation, we had 17 participants in which they evaluated, in average, 20 summaries and chose whether the automatically generated summary is "relevant" or "not relevant" to a given query.

As for the second evaluation, we use ROUGE-n metric (Recall-Oriented Understudy Gisting Evaluation) [16] for computing the coverage of the automatically generated summaries against the manually created summaries in terms of a contiguous sequence of words (n-grams). For instance, $n = 1$ represents the unigram "European", while $n = 2$ represents the bigram "European Union". The coverage ratio of the contextualized summaries and the manually generated summaries for the length n is computed as follows:

$$ROUGE_n = \frac{\sum_{s \in S} \sum_{w_n \in s} |match(w_n)|}{\sum_{s \in S} \sum_{w_n \in s} |(w_n)|} \tag{3}$$

where $|match(w_n)|$ is the total number of the $n - grams$, represented as w_n, that are part of the automatically generated summary and the manually generated summaries, i.e. the reference summaries S. Obviously, ROUGE-n is a recall metric between a candidate summary and a set of reference summaries. Our evaluation was performed over 20 queries, which generated approximately 110 summaries on average per query. The manually created summaries extracted from the NYT corpus were used as reference summaries. Note that, the automatically generated summary and its reference summary correspond to the same document in the corpus.

[10] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19`

6.3 Results

The first evaluation used manual evaluators and aimed at assessing the relevance of a summary given a user query, 76% of the automatically generated summaries were marked as "relevant".

For the second evaluation, we used ROUGE-1 to compare the automatically generated summaries and the manually generated abstractive summaries. Fig. 4 summarizes the results obtained by a sample of user queries. Our method achieved 25% precision for the query "Super Bowl", which is a comparatively high precision value for such task. Furthermore, the query "Terrorist Attacks" obtained 32% in terms of recall. The F_1 measure ranged from 12% to 26%, which is comparable to traditional summarization techniques.

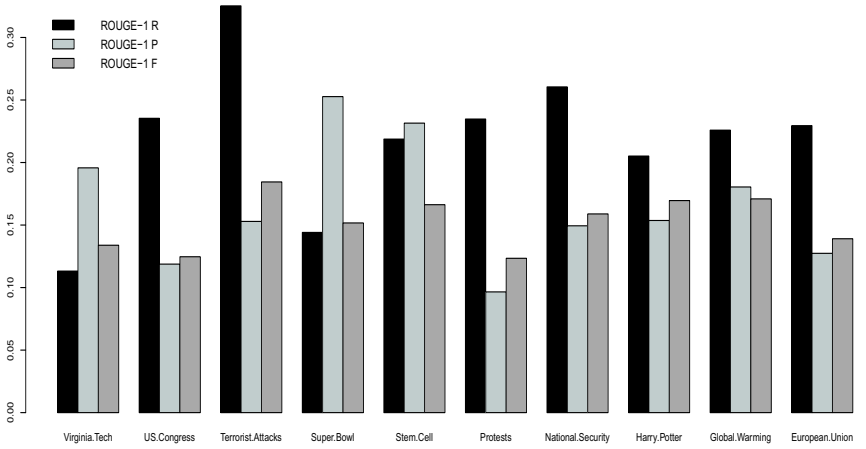

Fig. 4. Results for ROUGE-1 metric for different queries

Table 2. Generated focused summaries for different queries

Query	European Union	Super Bowl	US Congress	Virginia Tech	Stem Cell	Protest	Harry Potter	Global Warming	National Security	Terrorist Attacks
#Q.Terms	7	13	17	28	5	2	22	5	0	0
#Doc.	157	370	13	12	105	129	10	198	250	57
#Summ.	129	325	19	11	86	103	7	170	207	52

In Table 3 we show a small subset of generated summaries for the evaluation queries reported in Table 2. For readability reasons, we do not show all the information about **entities** and **actions**, and their contexts, however, we indicate the two different structures with colors in blue and red, respectively.

Table 3. Sample generated focused summaries from retrieved documents for the reported evaluation queries

Query: "European Union"	Query: "Super Bowl"
Bulgaria → joined → European Union European Union on Monday → helping to end → geographic divisions → left → cold war → extending → borders of the now 27-member bloc eastward to the Black Sea. Georges Prtre → is → former music director of the Paris Opera → has conducted → most world → leading → symphony orchestras.	New York Giants → are to realize → Super Bowl they → held so firmly → beginning of the season → felt completely → implausible weekend they → have to win → three games on the road. Philadelphia Eagles → have played → N.F.C. championship games in the past past years → reached → Super Bowl after the season → losing to → New England.
Query: "National Security'	Query: "Virginia Tech"
Kissinger Henry A (Dr) → was named → secretary of state in while → keeping → post as national security adviser. Republicans → forced to → Congressional sidelines for the first time in years → growing increasingly agitated → Democratic timetable.	Clemson University → try to start → new streak Wednesday University of Maryland → plays → host Carolina → lost to → unranked Virginia Tech on Saturday. Virginia → needed → mountain-sized comeback → topple → Georgia Tech in the Gator Bowl Louisville → took → advantage of some timely turnovers to → outlast → Wake Forest.
Query: "Stem Cell"	Query: "Protest"
Republicans → boasted → support for embryonic stem cell research as a way to → find → treatments for a wide range of diseases. Democrats → applauded → Mr. Spitzer Eliot (Gov) calls → insure → 500000 children → lack → health insurance → enroll → 900000 adults → are → eligible Medicaid → enrolled → issue debt → pay → stem cell research.	Students → clashed → police in this country last May attention → focused not just → demands → hold → elections without government meddling leaders → organizing → protests. Submarine → rammed → Japanese fishing vessel in waters off Hawaii → killing → nine people.
Query: "Global Warming"	Query: "Harry Potter"
Scientists over how to → describe → climate threat → is particularly → intense experts → work → final language in portions of the latest assessment of global warming by the Intergovernmental Panel on Climate Change. Scientists → shouting lately → global warming → is → human-caused catastrophe.	Dresden → played → Blackthorne Paul Blackthorne Paul → is → Harry Potter → grown up to become → Columbo. America → taking → children movies → has become → central cultural activity.
Query: "Terrorist Attacks"	Query: "US Congress"
Homeland Security Department → is essentially → first line of defense again terrorist attacks → is serving → nation. Pentagon → has increased → domestic intelligence collection efforts → help ensure → American bases → are protected → potential terrorist attacks.	Proposal → being considered → small businesses → allow write → larger part of they → go to → court → challenge → federal regulations. Bush George W (Pres) → has been → bit forthright things → have gone → Iraq Cheney Dick (Vice Pres) → spoke → enormous successes → refused to pay even → curled-lip service → consulting → Congress.

7 Conclusions and Future Work

Our approach addresses the task of focused knowledge extraction applied to the problem of generating focused entity-centric summaries for a given user query. We exploit POS pattern analysis and NER techniques to identify relevant statements and entities within a certain context to automatically generate query-based summaries. We also provide an RDF schema with the structured summaries for further reasoning in a publicly available knowledge base, which directly contributes to create a body of knowledge about entities and their appearance contexts over time. Furthermore, the techniques presented in this paper expand state of the art techniques on text summarization as well as information extraction.

We extensively evaluated our approach in order to validate that the automatically generated summaries address the user query needs and that it covers the main concepts of the documents. Indeed, our results showed that 76% of the summaries were relevant to the user queries and the concepts contained in the query. Moreover, our automatic evaluation proved to be comparable to state of the art techniques when assessed using the ROUGE-1 metric. In terms of the best performing queries, the results for precision, recall and F1 reached 25% of precision for the query "Super Bowl", 35% of recall and a F1 of 26% for the query "Terrorist Attacks". This shows that our approach extracted focused knowledge with high precision by incorporating the user interests through the query terms and it detected the importance of specific POS tags after a POS analysis of the terms in different topics.

As part of future work, we are working on reducing the number of patterns generated for a query. Since, it is a combinatorial problem when looking for patterns that involve many query terms. However, this problem could be circumvented by introducing a prior language analysis step to constrain the number of patterns that are appropriate. Moreover, we plan to apply this technique to several other domains.

References

1. Augenstein, I., Padó, S., Rudolph, S.: Lodifier: Generating linked data from unstructured text. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 210–224. Springer, Heidelberg (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Bouayad-Agha, N., Casamayor, G., Wanner, L., Díez, F., López Hernández, S.: FootbOWL: Using a generic ontology of football competition for planning match summaries. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 230–244. Springer, Heidelberg (2011)
4. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. Inf. Process. Manage. 31(5), 675–685 (1995)
5. Bryl, V., Giuliano, C., Serafini, L., Tymoshenko, K.: Supporting natural language processing with background knowledge: Coreference resolution case. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 80–95. Springer, Heidelberg (2010)
6. Cheng, G., Tran, T., Qu, Y.: Relin: Relatedness and informativeness-based centrality for entity summarization. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 114–129. Springer, Heidelberg (2011)
7. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: A framework and graphical development environment for robust nlp tools and applications. In: ACL, pp. 168–175 (2002)
8. Dietze, S., Maynard, D., Demidova, E., Risse, T., Peters, W., Doka, K., Stavrakas, Y.: Entity extraction and consolidation for social web content preservation. In: SDA, pp. 18–29 (2012)

9. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM 51(12), 68–74 (2008)
10. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: EMNLP, pp. 1535–1545 (2011)
11. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL (2005)
12. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: SIGIR, pp. 19–25 (2001)
13. Grefenstette, G.: Short query linguistic expansion techniques: Palliating one-word queries by providing intermediate structure to text. In: Pazienza, M.T. (ed.) SCIE 1997. LNCS, vol. 1299, pp. 97–114. Springer, Heidelberg (1997)
14. Hovy, D., Fan, J., Gliozzo, A.M., Patwardhan, S., Welty, C.A.: When did that happen? - linking events and relations to timestamps. In: EACL, pp. 185–193 (2012)
15. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, CONLL Shared Task 2011, Stroudsburg, PA, USA, pp. 28–34. Association for Computational Linguistics (2011)
16. Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Marie-Francine Moens, S.S. (ed.) Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics (2004)
17. Mausam, M., Schmitz, S., Soderland, R.: Bart, and O. Etzioni. Open language learning for information extraction. In: EMNLP-CoNLL, pp. 523–534 (2012)
18. Pereira Nunes, B., Kawase, R., Dietze, S., Taibi, D., Casanova, M.A., Nejdl, W.: Can entities be friends? In: Reggio, G., Astesiano, E., Tarlecki, A. (eds.) Abstract Data Types 1994 and COMPASS 1994. LNCS, vol. 906, pp. 45–57. Springer, Heidelberg (1995)
19. Radev, D.R., McKeown, K.: Generating natural language summaries from multiple on-line sources. Computational Linguistics 24(3), 469–500 (1998)
20. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.D.: A multi-pass sieve for coreference resolution. In: EMNLP, pp. 492–501 (2010)
21. Ritter, A., Mausam, Etzioni, O., Clark, S.: Open domain event extraction from twitter. In: KDD, pp. 1104–1112 (2012)
22. Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: SIGIR, pp. 2–10 (1998)
23. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, Stroudsburg, PA, USA, vol. 1, pp. 173–180. Association for Computational Linguistics (2003)
24. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, EMNLP 2000, Stroudsburg, PA, USA, vol. 13, pp. 63–70. Association for Computational Linguistics (2000)

25. Wan, X.: Topic analysis for topic-focused multi-document summarization. In: CIKM, pp. 1609–1612 (2009)
26. Wang, D., Zhu, S., Li, T., Chi, Y., Gong, Y.: Integrating document clustering and multidocument summarization. TKDD 5(3), 14 (2011)
27. White, M., Korelsky, T.: Multidocument summarization via information extraction. In: Proceedings of the HLT Conference, pp. 263–269 (2001)
28. Zhou, Y., Guo, Z., Ren, P., Yu, Y.: Applying wikipedia-based explicit semantic analysis for query-biased document summarization. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2010. LNCS, vol. 6215, pp. 474–481. Springer, Heidelberg (2010)