Multi-document summarization based on the Yago ontology

Multi-document summarization based on the Yago ontology

Elena Baralis, Luca Cagliero*, Saima Jabeen

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Alessandro Fiori

IRC@C: Institute for Cancer Research at Candiolo, Str. Prov. 142 Km. 3.95 10060 - Candiolo (TO) - Italy

Sajid Shah

Dipartmento di Elettronica e Telecomunicazioni, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Abstract

Sentence-based multi-document summarization is the task of generating a succinct summary of a document collection, which consists of the most salient document sentences. In recent years, the increasing availability of semantics-based models (e.g., ontologies and taxonomies) has prompted researchers to investigate their usefulness for improving summarizer performance. However, semantics-based document analysis is often applied as a preprocessing step, rather than integrating the discovered knowledge into the summarization process.

This paper proposes a novel summarizer, namely Yago-based Summarizer,

*Corresponding author. Tel.: +39 011 090 7084. Fax: +39 011 090 7099.

Email addresses: elena.baralis@polito.it (Elena Baralis), luca.cagliero@polito.it (Luca Cagliero), saima.jabeen@polito.it (Saima Jabeen), alessandro.fiori@ircc.it (Alessandro Fiori), sajid.shah@polito.it (Sajid Shah)

Preprint submitted to

that relies on an ontology-based evaluation and selection of the document sentences. To capture the actual meaning and context of the document sentences and generate sound document summaries, an established entity recognition and disambiguation step based on the Yago ontology is integrated into the summarization process.

The experimental results, which were achieved on the DUC'04 benchmark collections, demonstrate the effectiveness of the proposed approach compared to a large number of competitors as well as the qualitative soundness of the generated summaries.
Keywords: Document Summarization, Text Mining, Entity Recognition

1. Introduction

Discovering the most salient information hidden in textual Web documents is often a challenging task. In fact, the huge volume of electronic documents that users could retrieve from the Web is commonly very difficult to explore without the help of automatic or semi-automatic tools. To tackle this issue, a particular attention has been paid to the development of text summarization tools. Summarizers focus on generating a succinct representation of a textual document collection. Specifically, sentence-based multi-document summarizers generate concise yet informative summaries of potentially large document collections, which consist of the most representative document sentences.

A significant research effort has been devoted to tackling the summarization problem by means of general-purpose information retrieval or data mining techniques. For example, clustering-based approaches (e.g., [47, 48])

adopt clustering algorithms to group document sentences into homogeneous clusters and then select the most authoritative representatives within each group. In contrast, graph-based approaches (e.g., [36, 51, 52]) first generate a graph-based model in which the similarity relationships between pairs of sentences are represented. Next, they exploit popular indexing strategies (e.g., PageRank [9]) to identify the most salient sentences (i.e., the most authoritative graph nodes). However, in some cases, the soundness and readability of the generated summaries are unsatisfactory, because the summaries do not cover in an effective way all the semantically relevant data facets. A step beyond towards the generation of more accurate summaries has been made by semantics-based summarizers (e.g., [12, 13]). Such approaches combine the use of general-purpose summarization strategies with ad-hoc linguistic analysis. The key idea is to also consider the semantics behind the document content to overcome the limitations of general-purpose strategies in differentiating between sentences based on their actual meaning and context.

Ontologies are formal representations of the most peculiar concepts that are related to a specific knowledge domain and their corresponding relationships [5]. Ontologies find application in several research contexts, among which user-generated content analysis [20], e-learning platform development [25], and video and image analysis [45]. In recent years, the attention of the research community has been focused on both learning meaningful ontologies that contain salient document keywords [8] and improving the performance of the document summarization process by integrating ontological knowledge [21, 24, 34, 35]. For example, ontologies have been used to identify the document concepts that are strongly correlated with a user-specified

query [24, 34] or to map the document content to non-ambiguous ontological concepts [21, 35]. However, most the previously proposed approaches perform the semantics-based analysis as a preprocessing step that precedes the main summarization process. Therefore, the generated summaries could not entirely reflect the actual meaning and context of the key document sentences. In contrast, we aim at tightly integrating the ontology-based document analysis into the summarization process in order to take the semantic meaning of the document content into account during the sentence evaluation and selection processes. With this in mind, we propose a new multi-document summarizer, namely Yago-based Summarizer, that integrates an established ontology-based entity recognition and disambiguation step. Specifically, a popular ontological knowledge base, i.e., Yago [41], is used to identify the key document concepts. The same concepts are also evaluated in terms of their significance with respect to the actual document context. The result of the evaluation process is then used to select the most representative document sentences. In such a way, the knowledge that is inferred from Yago is tightly integrated into the sentence evaluation process. Finally, a variant of the Maximal Marginal Relevance (MMR) evaluation strategy [10] is adopted to iteratively choose the best subset of informative yet non-redundant sentences according to the previously assigned sentence ranks.

To demonstrate the effectiveness the proposed approach, we compared its performance on the DUC'04 benchmark document collections with that of a large number of state-of-the-art summarizers. Furthermore, we also performed a qualitative evaluation of the soundness and readability of the generated summaries and a comparison with the results that were produced

by the most effective summarizers.

This paper is organized as follows. Section 2 compares our approach with the most recent related works. Section 3 presents and thoroughly describes the Yago-based Summarizer system. Section 4 experimentally evaluates the effectiveness and usefulness of the proposed approach, whereas Section 5 draws conclusions and presents future developments of this work.

2. Related work

A significant research effort has been devoted to summarizing document collections by exploiting information retrieval or data mining techniques. Two main summarization strategies have been proposed in literature. *Sentence-based* summarization focuses on partitioning documents in sentences and generating a summary that consists of the subset of most informative sentences (e.g., [11, 29, 47]). In contrast, *keyword-based* approaches focus on detecting salient document keywords using, for instance, graph-based indexing [28, 51, 52] or latent semantic analysis [16]. Since sentence-based approaches commonly generate humanly readable summaries without the need for advanced postprocessing steps, our summarizer relies on a sentence-based approach. Summarizers can be further classified as constraint-driven if they entail generating a summary that satisfy a set of (user-specified) constraints [2]. For example, sentences that are pertinent to a user-specified query can be selected [30, 33]. Unlike [2, 30, 33] our summarizer relies on a constraint-less approach.

Most of the recently proposed (constraint-less) sentence-based summarizers exploit one of the following general-purpose techniques: (i) clustering, (ii)

graph mining, (iii) linear programming, and (iv) itemset mining. Clustering-based approaches (e.g., [47, 48]) group document sentences into homogeneous clusters and then select the best representatives (e.g., the centroids or the medoids [32]) within each cluster. While the authors in [47] propose a static summarization framework, the work that was first presented in [48] addresses the problem of incremental summary update: whenever a set of documents is added/removed from the initial collection, the previously generated summary is updated without the need for recomputing the whole clustering model. In parallel, some attempts to cluster documents rather than sentences have also been made [37, 7]. For example, MEAD [37] analyzes the cluster centroids and generates a pseudo-document that includes the sentences with the highest tf-idf term values [27]. Then, the sentence selection process is driven by a score that considers (i) the sentence similarity with the centroids, (ii) the sentence position within the document, and (iii) the sentence length. A similar approach has also been adopted to summarize articles coming from the biological domain [7]. To tailor the generated summaries to the most relevant biological knowledge biologists are asked to provide a dictionary that is used to drive the sentence selection process. Similarly, this work also considers the document context to improve the summarization performance. Unlike [7], it exploits an ontological knowledge base, rather than a plain-text dictionary, to drive the sentence evaluation and selection process.

Graph-based approaches to sentence-based summarization (e.g., [36, 44, 46, 51, 52]) generate a graph in which the nodes represent the document sentences, whereas the edges are weighted by a similarity measure that is evaluated on each node pair. Popular indexing strategies (e.g., PageRank [9],

HITS [23]) are exploited to rank the sentences based on their relative authoritativeness in the generated graph. In parallel, other approaches formalize the sentence selection task as a min-max optimization problem and tackle it by means of linear programming techniques [1, 3, 17, 42]. Still others analyze the underlying correlations among document terms by exploiting (i) frequent itemset mining techniques [6], (ii) probabilistic approaches [12, 13], or (iii) the Singular Value Decomposition (SVD) [40].

Ontologies have already been exploited to improve the document summarization performance. Specifically, they have been used to (i) identify the concepts that are either most pertinent to a user-specified query [24, 34] or most suitable for performing query expansion [31], (ii) model the context in which summaries are generated in different application domains (e.g., the context-aware mobile domain [18], the business domain [50], the disaster management domain [26]), and (iii) enrich existent ontological models with textual content [8]. Some attempts to consider the text argumentative structure into account during the summarization process have also been made. For example, in [35] the authors propose to identify and exploit salient lexical chains to generate accurate document summaries. The summarizer proposed in [21] exploits Support Vector Machines (SVMs) to maps each sentence to a subset of taxonomy nodes. Similarly, in [4] a rhetorical role is assigned to each sentence by a stochastic CRF classifier [39], which is trained from a collection of annotated sentences. Unlike [4, 21] our approach does not rely on classification models. Furthermore, since our summarizer evaluates the sentence relevance regardless of the underlying document structure, our approach is, to some extent, complementary to the ones that have previously

been proposed in [4, 35].

3. Yago-based Summarizer

Yago-based Summarizer is a novel multiple-document summarizer that exploits the Yago ontological knowledge base [41] to generate accurate document summaries.

Consider a collection of textual documents $D=\{d_1, \ldots, d_N\}$, where each document $d_i \in D$ is composed of a set of sentences s_1^i, ..., s_M^i. The summarizer generates a summary $S=\{s_j^i\}$ $1 \leq i \leq N$, $1 \leq j \leq M$. The summary includes a worthwhile subset of sentences that are representative of the whole collection D.

Figure 1 outlines the main Yago-based Summarizer steps, which are briefly summarized below.

- **Entity recognition and disambiguation.** This step analyzes the input document collection with the goal of identifying the most relevant concepts and their corresponding context of use. To this aim, the Yago knowledge base is used to map the words that occur in the document sentences to non-ambiguous ontological concepts, called *entities*. To discriminate between multiple candidate entities for the same word combination, it adopts an entity relevance score that considers both the popularity and the contextual pertinence of each candidate entity in the analyzed document collection.

- **Sentence ranking.** To include in the summary only the most pertinent and semantically meaningful document content, sentences are evaluated and ranked according to the previously assigned entity scores.

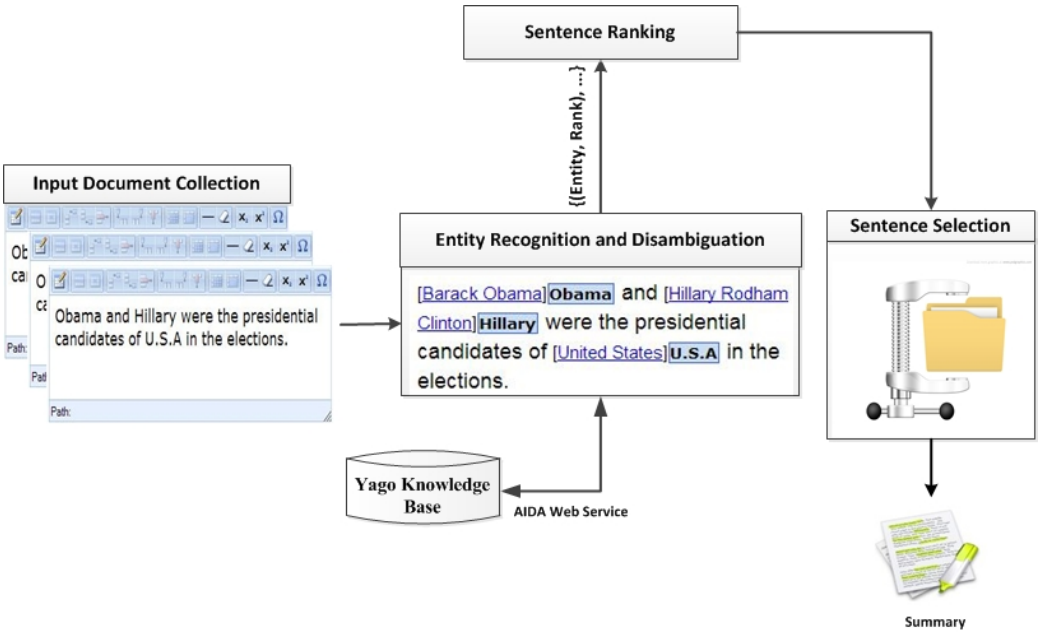


Figure 1: The Yago-based Summarizer.

- **Sentence selection.** To generate a summary of the document collection an iterative procedure is applied to select the top-ranked sentences that are least similar to the previously selected ones.

3.1. Entity recognition and disambiguation

Entity recognition and disambiguation are established document analysis tasks that aim at mapping the natural text to a set of non-ambiguous ontological concepts [32]. Yago-based Summarizer exploits the Yago ontological knowledge base [41], which relies on the Wikipedia free encyclopedia [49], to support the entity recognition and disambiguation process. The Yago analytical procedures have been called through the AIDA Web Service [22].

Consider a sentence s_j^i that is composed of a collection of (possibly repeated) words w_1, w_2, ..., w_Z. The goal is to map words w_k, $1 \leq k \leq Z$

to Yago ontological concepts, i.e., the entities. Note that an entity may be associated either with a single word or with a combination of words. The entity recognition step recognizes the entities that are associated with noun, dates, times, or numbers. As a clarifying example, consider the sentence reported in the left-hand side of Figure 2.

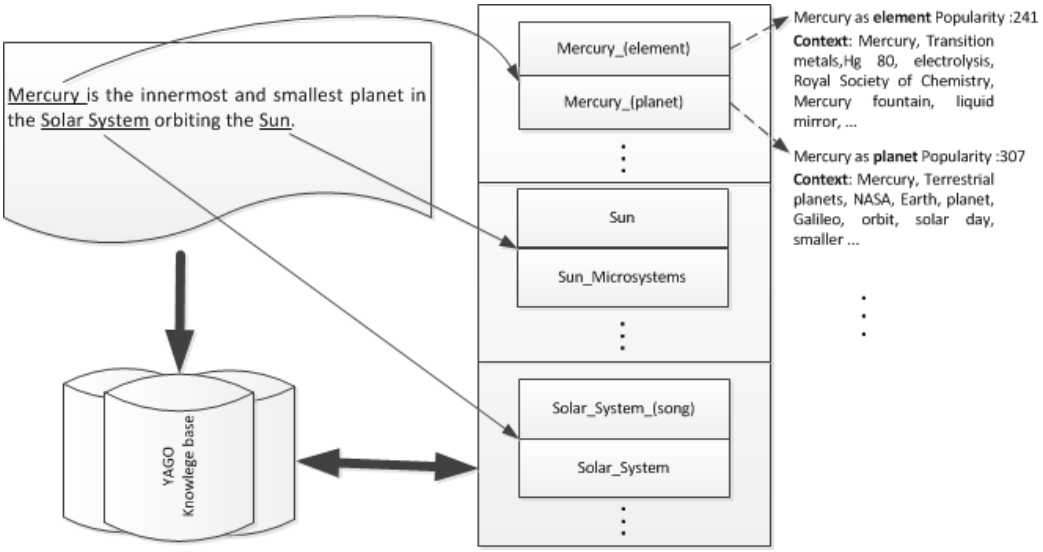


Figure 2: Entity Recognition and Disambiguation example.

Each of the three underlined word combinations *Mercury*, *Solar System*, and *Sun* is associated with at least one candidate entity in Yago. Note that not all the sentence words match at least one Yago entity. For example, the word *Innermost* has no matching entity. Furthermore, some words have many candidate entities, meaning that a word could have different meanings in different contexts. For example, *Mercury* could be associated with the candidate entities *Mercury(Element)* and *Mercury(Planet)*, which correspond to the well-known chemical element and planet, respectively. Note also that for each entity Yago provides (i) a popularity score, which reflects

its frequency of usage (e.g., 241 for *Mercury(Element)*), (ii) a list of related keywords (e.g., *Chemistry*, *Liquid*) for its corresponding context of use, and (iii) the number of incoming and outcoming Wikipedia links [49].

Entity recognition for times, date, and numbers is based on regular expressions and returns a single entity. For example, the expression *March, 1st 2012* corresponds to the date 01/03/2012. Conversely, the entity recognition procedure for nouns could return many candidate entities. Hence, in the latter case a disambiguation step is applied in order to select, among the candidate entities, the most appropriate one. To tackle this issue, each candidate entity is weighted by a relevance score, which considers both its popularity and pertinence to the analyzed document context. Specifically, the rank $entityRank(e_q)$ of an entity e_q with respect to a word w_k that occurs in the document $d_i \in D$, is defined as follows:

$$\begin{aligned} entityRank(e_q) &= \theta \cdot popularity(e_q) \\ &+ \phi \cdot sim(cxt(e_q), cxt(d_i)) \\ &+ (1 - \theta - \phi) \cdot coh(e_q, D) \end{aligned} \tag{1}$$

where $\theta, \phi \in [0, 1]$ are user-specified parameters that weigh the importance of each summation term, $popularity(e_q)$ is the Yago popularity score that is associated with the candidate entity e_q, $sim(cxt(e_q),cxt(d_i))$ is the similarity between the context of use of the candidate entity and the document d_i, and $coh(e_q, D)$ is the coherence of e_q with respect to the whole document collection. By following the indications reported in [22], we set the values of θ and ϕ to 0.34 and 0.47, respectively. A thorough assessment of the

entity recognition system performance on real data is also reported in [19]. The first summation term is a popularity score, which indicates the global frequency of occurrence of the concept in the knowledge base. For instance, in Yago *Mercury-Planet* has, on average, a higher popularity score than *Mercury-Element* (307 against 241). However, the relevance of an entity within a document also depends on its context of use. Hence, the second summation term indicates the pertinence of the entity to the document. Specifically, it measures the cosine distance [32] between the context of the candidate entity e_q, i.e., the list of contextual keywords that are provided by Yago (e.g., *Chemistry*, *Liquid* for the candidate entity *Mercury-Element* in Figure 2), and the context of the word w_k at the document level (i.e., the list of words that co-occur with w_k in d_i). Roughly speaking, the more contextual keywords match the document content the higher the pertinence of the candidate entity is. Finally, since the recognized entities are likely to be correlated each other, the last summation term measures the coherence of the candidate entity with respect to all of the other recognized candidate entities that correspond to any word in D. Since coherent entities are likely to share many Wikipedia links, similar to [22], we evaluate the entity coherence within the document collection D as the number of incoming Wikipedia links that are shared by e_q and all of the other candidate entities that have been recognized in D.

The entity scores will be used to drive the summary generation process, as discussed in the following sections.

3.2. Sentence ranking

Yago-based Summarizer exploits the semantic knowledge that has been

inferred at the previous step to evaluate and rank the document sentences according to their significance in the document collection. To this aim, a rank is associated with each document sentence. The sentence rank reflects the relevance of the entities associated with its corresponding sentence words.

Let s_j^i be an arbitrary sentence and $E(s_j^i)$ the set of entities (nouns, date, times, or numbers) that are associated with any word $w_k \in s_j^i$. The s_j^i's rank is computed as follows:

$$SR(s_j^i) = \frac{\sum_{e_q \in E(s_j^i)} EntityScore(e_q)}{|E(s_j^i)|} \tag{2}$$

where $EntityScore(e_q)$ is defined by

$$EntityScore(e_q) = \begin{cases} \gamma & \text{if } e_q \text{ is a date, time, or number entity,} \\ \gamma + EntityRank(e_q) & \text{if } e_q \text{ is a named entity} \end{cases} \tag{3}$$

γ is a user-specified parameter that is used to privilege the sentences that contain many recognized entities. $\sum_{e_q \in E(s_j^i)} EntityScore(e_q)$ is the summation of the entity ranks of all of the entities that are mapped to any word in s_j^i (see Definition 1). Note that the sentences that do not contain any recognized Yago entity have minimal sentence rank (i.e., 0), because they are not likely to contain any semantically relevant concept. In contrast, because of the γ correction, the sentences that contain only dates, times, or numbers are considered to be, on average, more relevant than the former ones, but less relevant than those that also contain named entities. The impact of the user-specified parameter γ on the summarization performance is discussed in Section 4.

3.3. Sentence selection

Given a sentence ranking, the selection step focuses on generating the output summary of the document collection by including only the most representative sentences. To achieve this goal, Yago-based Summarizer adopts a variant of an established iterative re-ranking strategy, called Maximal Marginal Relevance (MMR) [10]. MRR has first been introduced in the context of query-based summary generation. At each algorithm iteration, it picks out the candidate sentence that is characterized by (i) maximal relevance with respect to the given query and (ii) minimal similarity with respect to the previously selected sentences. Since our approach is not query-based, we adapt the former selection strategy to the problem under analysis. Specifically, Yago-based Summarizer selects, at each iteration, the top-ranked sentence with minimal redundancy with respect to the already selected sentences. At each iteration the former optimization problem can be formulated as follows:

$$\begin{aligned} &\underset{\{s_j^i\}}{\text{maximize}} \quad \alpha \cdot SR(s_j^i) - (1-\alpha) \cdot sim(s_j^i, \bar{s}_t^r) \\ &\text{subject to} \\ &\qquad\qquad\qquad\qquad\qquad\qquad \alpha \in [0,1] \\ &\qquad\qquad\qquad\qquad\qquad\qquad s_j^i \notin S \\ &\qquad\qquad\qquad\qquad\qquad\qquad \bar{s}_t^r \in S \end{aligned} \tag{4}$$

where S is the output summary that possibly includes some of the document sentences $\bar{s}_t^r$, α is a user-specified parameter, and $\{s_j^i\}$ is the set of candidate sentences not yet included in the summary. The sentence ranking is evaluated using the expression reported in Formula 2. Furthermore, the similarity

$sim(s_j^i, \bar{s}_t^r)$ between the pair of sentences s_j^i and $\bar{s}_t^r$ is evaluated using the cosine similarity [32] and takes value zero when the summary is empty (i.e., at the first algorithm iteration). The impact of the entity relevance and the similarity score is weighted by the α parameter. Specifically, the higher the value of α is, the more important the entity relevance score is with respect to the similarity score. Therefore, setting relatively high α values could yield informative but partially redundant summaries. Conversely, for lower α values the sentence relevance is partially neglected in behalf of a lower summary redundancy.

4. Experimental results

We performed a variety of experiments to address the following issues: (i) a performance comparison between Yago-based Summarizer and many state-of-the-art summarizers on document benchmark collections (see Section 4.2), (ii) a qualitative comparison between the summaries generated by our approach and those produced by two representative competitors (see Section 4.3), and (iii) an analysis of the impact of the main system parameters on the Yago-based Summarizer performance (see Section 4.4).

All the experiments were performed on a 3.0 GHz 64 bit Intel Xeon PC with 4 GB main memory running Ubuntu 10.04 LTS (kernel 2.6.32-31). The source code for Yago-based Summarizer is available, for research purposes, upon request to the authors. A detailed description of the experimental evaluation context is given below.

4.1. Evaluation context

We evaluated the Yago-based Summarizer performance on the task 2 of the Document Understanding Conference (DUC) 2004, which is the latest benchmark contest that were designed for generic English-written multi-document summarization [47]. The analyzed DUC'04 collections have been provided by the contest organizers [15]. They consist of a large variety of English-written articles which range over different subjects. According to their subject, articles were preliminary clustered in 50 document groups. Each homogeneous collection contains approximately 10 documents. Furthermore, for each collection at least one *golden* summary is given by the DUC'04 organizers. Participants to the DUC'04 contest had to submit their own summaries and compare them with the reference (golden) ones. The more similar the generated summaries are to the reference models, the more accurate the summarization process is.

To perform an analytical comparison between the summarizers' performance on the task 2 of DUC'04 we used the ROUGE toolkit [27], which has been adopted as official DUC'04 tool for performance evaluation[1]. ROUGE measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries (i.e., the golden summaries). The summary that achieves the highest ROUGE score could be considered to be the most similar to the golden summary. To perform a fair comparison, before using the ROUGE toolkit we normalized the generated summaries by truncating each of them at 665 bytes (we round the number

[1]The provided command is: ROUGE-1.5.5.pl -e data -x -m -2 4 -u -c 95 -r 1000 -n 4 -f A -p 0.5 -t 0 -d -a

down in case of straddled words). Several automatic evaluation scores are implemented in ROUGE. As previously done in [6, 47], we will report only the ROUGE-2 and ROUGE-4 representative scores [27]. Similar results were achieved for the other ROUGE scores.

4.2. Performance comparison on the DUC'04 collections

We compared the Yago-based Summarizer performance on the DUC'04 benchmark collections with that of: (i) the 35 summarizers submitted to the DUC'04 conference, (ii) the 8 summaries generated by humans and provided by the DUC'04 system (beyond the golden summaries), (iii) two widely used open source text summarizers, i.e., the Open Text Summarizer (OTS) [38] and TexLexAn [43], (iv) a recently proposed itemset-based summarizer [6], named ItemSum (Itemset-based Summarizer), and (v) a baseline version of Yago-based Summarizer, namely Baseline, which adopts an established term relevance evaluator, i.e., the tf-idf score [27], rather than the ontology-based entity rank evaluator (see Definition 1).

For the DUC'04 competitors we considered the results that were provided by the DUC'04 system [15]. Specifically, for the top-ranked DUC'04 summarizer, i.e., CLASSY [13], we considered its most effective version (i.e., peer65). Similarly, for the other competitors we tuned the algorithm parameters to their average best value by following the indications that were given by the respective authors. For Yago-based Summarizer we set, as *standard* configuration, γ to 0.3 and α to 0.9. In Section 4.4 we analyze more in detail the impact of both parameters on the Yago-based Summarizer performance.

Table 1 summarizes the results that were achieved by Yago-based Summarizer, Baseline-tf-idf, ItemSum, OTS, TexLexAn, the 8 humanly generated

summaries, and the 10 most effective summarizers presented in the DUC'04 contest. To validate the statistical significance of the Yago-based Summarizer performance improvement against its competitors we performed the paired t-test [14] at 95% significance level for all of the evaluated measures. Every statistically relevant worsening in the comparison between Yago-based Summarizer and the other approaches is starred in Table 1.

Table 1: DUC'04 Collections. Comparisons between Yago-based Summarizer and the other approaches. Statistically relevant differences in the comparisons between Yago-based Summarizer (standard configuration) and the other approaches are starred.

Summarizer		ROUGE-2			ROUGE-4		
		R	Pr	F	R	Pr	F
TOP RANKED DUC'04 PEERS	peer120	0.076*	**0.103**	0.086*	0.014*	**0.019**	0.016
	peer65	0.091*	0.090*	0.091*	0.015*	0.015	0.015*
	peer19	0.080*	0.080*	0.080*	0.010*	0.010*	0.010*
	peer121	0.071*	0.085*	0.077*	0.012*	0.014*	0.013*
	peer11	0.070*	0.087*	0.077*	0.012*	0.015*	0.012*
	peer44	0.075*	0.080*	0.078*	0.012*	0.013*	0.012*
	peer81	0.077*	0.080*	0.078*	0.012*	0.012*	0.012*
	peer104	0.086*	0.084*	0.085*	0.011*	0.010*	0.010*
	peer124	0.083*	0.081*	0.082*	0.012*	0.012*	0.012*
	peer35	0.083*	0.084	0.083*	0.010*	0.011*	0.011*
DUC'04 HUMANS	A	0.088*	0.092*	0.090*	0.009*	0.010*	0.010*
	B	0.091*	0.096	0.092	0.013*	0.013*	0.013*
	C	0.094	0.102	0.098	0.011*	0.012*	0.012*
	D	0.100	**0.106**	0.102	0.010*	0.010*	0.010*
	E	0.094	0.099	0.097	0.011*	0.012*	0.012*
	F	0.086*	0.090*	0.088*	0.008*	0.009*	0.009*
	G	0.082*	0.087*	0.084*	0.008*	0.008*	0.007*
	H	**0.101**	0.105	**0.103**	0.012*	0.013*	0.012*
OTS		0.075*	0.074*	0.074*	0.009*	0.009*	0.009*
texLexAn		0.067*	0.067*	0.067*	0.007*	0.007*	0.007*
ItemSum		0.083*	0.085*	0.084*	0.012*	0.014*	0.014*
Baseline		0.092*	0.091*	0.092*	0.014*	0.014*	0.014*
Yago-based Summarizer		**0.095**	0.094	**0.095**	**0.017**	0.017	**0.017**

Yago-based Summarizer performs significantly better than ItemSum, OTS, TexLexAn, and Baseline for all of the analyzed measures. Hence, the ontology-based sentence ranking and selection strategies appear to be more effective than traditional information retrieval techniques (e.g., the tf-idf-based sentence evaluation [27]) for summarization purposes. Although, in some cases,

peer120 performs best in terms of ROUGE-2 and ROUGE-4 precision, Yago-based Summarizer performs significantly better than all the 35 DUC'04 competitors in terms of ROUGE-2 and ROUGE-4 F1-measure (i.e., the harmonic average between precision and recall). Hence, the summaries that were generated by Yago-based Summarizer are, on average, the most accurate and not redundant ones.

Compared to the 8 humanly generated summaries, Yago-based Summarizer significantly outperforms 3 out of 8 and 8 out of 8 competitors in terms of ROUGE-2 and ROUGE-4 F1-measure, respectively. In contrast, CLASSY (peer65) performs significantly better than 2 out of 8 and 6 out of 8 humans in terms of ROUGE-2 and ROUGE-4 F1-measure, respectively. Similarly, peer120 performs worser than all the humans in terms of ROUGE-2 and outperforms 7 out of 8 humans in terms of ROUGE-4. Hence, Yago-based Summarizer placed, on average, better than CLASSY and peer120 with respect to the humans.

4.3. Summary comparison

We conducted a qualitative evaluation of the soundness and readability of the summaries that were generated by Yago-based Summarizer and the other approaches. Tables 2 reports the summaries that were produced by Yago-based Summarizer, the top-ranked DUC'04 summarizer, i.e., CLASSY [13] (Peer-65), and a commonly used open source summarizer OTS [38] on a representative DUC'04 collection, which relates the activities and the main achievements of the Yugoslav war crime tribunal.

The summary that was generated by Yago-based Summarizer appears to be the most focused one, because it covers all the main document topics,

Table 2: Summary examples.

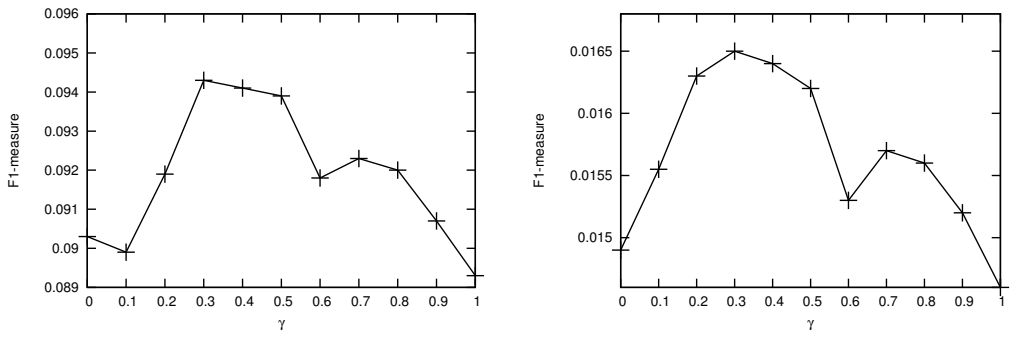
Method	**Summary**
Yago-based Summarizer	Yugoslavia must cooperate with the U.N. war crimes tribunal investigating alleged atrocities during the wars in Croatia and Bosnia, international legal experts meeting in Belgrade said Sunday. The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities. American and allied forces in Bosnia on Wednesday arrested a Bosnian Serb general who was charged with genocide by the international war crimes tribunal in a recent secret indictment.
CLASSY	Some of the closest combat in the half year of the Kosovo conflict, to the point of fighting room to room and floor to floor, occurred near this village six weeks ago, in the days before 21 women, children and elderly members of the Delijaj clan were massacred by Serbian forces, their mutilated bodies left strewn on the forest floor. In its first case to deal with atrocities against Serbs during Bosnia's civil war, a U.N. war crimes tribunal on Monday convicted three prison officials and guards, but acquitted a top military commander who oversaw the facility. Hundreds of people gathered at Sarajevo airport on Saturday to welcome Zejnil Delalic, who was cleared of war crimes charges earlier this week after spending 980 days in jail of the international war crimes tribunal in The Hague.
OTS	The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities. The Yugoslav war crimes tribunal cleared Zejnil Delalic, a Muslim, of responsibility for war crimes committed against Serb captives at a Bosnian government-run prison camp under his command. court convicted camp commander Zdravko Mucic, a Croat, of 11 war crimes and grave breaches of the Geneva Conventions because he oversaw guards who murdered nine Serbs and tortured six. Indeed, the conflict between Serbian forces bent on keeping Kosovo in Serbia and guerrillas fighting for the independence of its heavily ethnic Albanian population first drew international attention with the massacre of the Jasari clan in early March by Serbian units at Prekaz, in central Kosovo.

i.e., (1) the role of the Yugoslav war crime tribunal, (2) the acquittal of the Muslim military commander, and (3) the arrest of the Bosnian Serb general. In contrast, OTS and CLASSY cover, to some extent, only the topic (2). On the other hand, both OTS and CLASSY select other contextual sentences about the Kosovo war, which are very general and not representative of the key document message. Hence, the corresponding summaries are deemed to be partially redundant.

4.4. Parameter setting

The setting of the user-specified α and γ parameters could affect the Yago-based Summarizer performance significantly. Hence, we thoroughly analyzed their impact on the Yago-based Summarizer ROUGE scores.

Figures 3(a) and 3(b) plot the ROUGE-2 and ROUGE-4 F1-measure that were achieved by Yago-based Summarizer on the DUC'04 collection by varying the value of γ in the range [0,1] and by setting α to its best value (0.9), respectively. In contrast, Figures 4(a) and 4(b) plot the ROUGE-2 and ROUGE-4 F1-measure scores by setting the best γ value (0.3) and by varying α in the range [0,1].

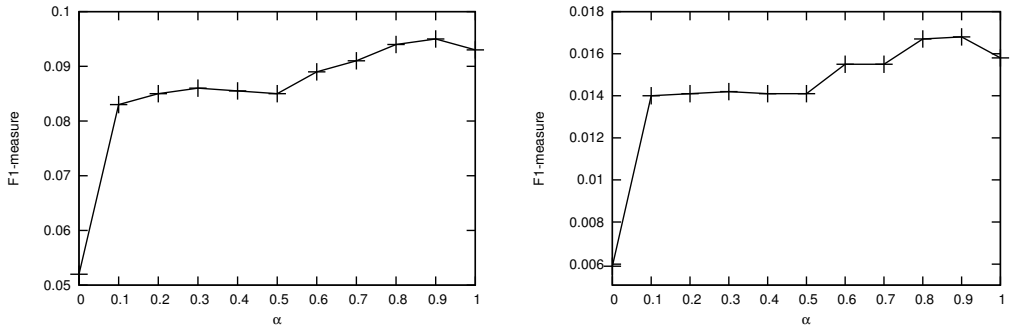


(a) ROUGE-2 F1-measure.

(b) ROUGE-4 F1-measure.

Figure 3: Impact of γ on the Yago-based Summarizer performance. α=0.9. DUC'04 collections.

When increasing the value of the γ parameter, the sentences that include many unrecognized words are on average penalized (see Formula 1). Based on the results that are reported in Figure 3, the best performance results were achieved by setting γ=0.3. Furthermore, the results remain relatively stable when γ ranges between 0.2 and 0.5. Since the EntityRank score of



(a) ROUGE-2 F1-measure.

(b) ROUGE-4 F1-measure.

Figure 4: Impact of α on the Yago-based Summarizer performance. γ=0.3. DUC'04 collections.

many of the recognized entities fall in the same value range, it means that redoubling the score of the recognized named entities with respect to the time/date/number entities yields good summarization performance. In contrast, setting γ out of the above value range implies giving an under- or over-emphasis to the least interesting entities.

The α parameter allows the user to decide to which extent the similarity between the already selected sentence is relevant compared to the entity-based sentence rank for sentence selection. The higher the value of α is, the more important the ontology-based sentence rank becomes with respect to the similarity with the previously selected sentences (see Formula 4). Since the analyzed documents contain a limited amount of redundancy, Yago-based Summarizer achieves averagely high ROUGE scores by setting high α values (i.e., $\alpha > 0.7$). With the DUC'04 collections, the best performance results were achieved by setting α=0.9. Note that, with such configuration setting, Yago-based Summarizer disregards, to a large extent, the impact of the sim-

ilarity score with respect to the ontology-based sentence ranking. However, when coping with document collections that contain a larger number of repetitions, the user should set lower α values in order to achieve a good trade-off between summary relevance and redundancy.

5. Conclusions and future works

In recent years, semantics-based document analysis has shown to improve the performance of document summarization systems significantly. However, since most of the related approaches perform semantics-based analysis as a preprocessing step rather than integrating the ontological knowledge into the summarization process, the quality of the generated summaries remains, in some cases, unsatisfactory.

This paper proposes to improve the performance of state-of-the-art summarizers by integrating an ontology-based sentence evaluation and selection step into the summarization process. Specifically, an established entity recognition and disambiguation step based on the Yago ontology is used to identify the key document concepts and evaluate their significance with respect to the document context. The same results are then exploited to select the most representative document sentences.

The experimental results show that Yago-based Summarizer performs better than many state-of-the-art summarizers on benchmark collections. Furthermore, a qualitative comparison between the summaries generated by Yago-based Summarizer and the state-of-the-art summarizers demonstrate the usefulness and applicability of the proposed approach.

Future developments on this work will address the application of the

proposed summarization approach to multilingual document collections and the use of entropy-based sentence selection strategies to further improve the compactness of the generated summaries.

References

[1] Alguliev, R. M., Aliguliyev, R. M., & Hajirahimova, M. S. (2012). Gendocsummclr: Generic document summarization based on maximum coverage and less redundancy. *Expert Systems with Applications*, *39*, 12460 – 12473. URL: http://www.sciencedirect.com/science/article/pii/S0957417412006641. doi:10.1016/j.eswa.2012.04.067.

[2] Alguliev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2013). Cdds: Constraint-driven document summarization models. *Expert Systems with Applications*, *40*, 458 – 465. URL: http://www.sciencedirect.com/science/article/pii/S0957417412009049. doi:10.1016/j.eswa.2012.07.049.

[3] Alguliev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications*, *40*, 1675 – 1689. doi:10.1016/j.eswa.2012.09.014.

[4] Atkinson, J., & Munoz, R. (2013). Rhetorics-based multi-document summarization. *Expert Systems with Applications*, *40*, 4346 4352. URL: http://www.sciencedirect.com/science/article/pii/S0957417413000304. doi:10.1016/j.eswa.2013.01.017.

[5] Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (Eds.) (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

[6] Baralis, E., Cagliero, L., Fiori, A., & Jabeen, S. (2012). Multi-document summarization exploiting frequent itemsets. In *In Proceedings of the ACM Symposium on Applied Computing (SAC 2012)*.

[7] Baralis, E., & Fiori, A. (2010). Summarizing biological literature with biosumm. In *CIKM* (pp. 1961–1962).

[8] Baxter, D., Klimt, B., Grobelnik, M., Schneider, D., Witbrock, M., & Mladenic, D. (2009). Capturing document semantics for ontology generation and document summarization. *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*, (pp. 141–154).

[9] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7* (pp. 107–117).

[10] Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* SIGIR '98 (pp. 335–336). New York, NY, USA: ACM. URL: `http://doi.acm.org/10.1145/290941.291025`. doi:10.1145/290941.291025.

[11] Carenini, G., Ng, R. T., & Zhou, X. (2007). Summarizing email conversations with clue words. In *World Wide Web Conference Series* (pp. 91–100).

[12] Conroy, J., Schlesinger, J., Kubina, J., Rankel, P., & OLeary, D. (2011). Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics. In *TAC'11: Proceedings of the The 2011 Text Analysis Conference*.

[13] Conroy, J. M., Schlesinger, J. D., Goldstein, J., & O'Leary, D. P. (2004). Left-Brain/Right-Brain Multi-Document Summarization. In *DUC 2004 Conference Proceedings*.

[14] Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *10*.

[15] Document Understanding Conference (2004). HTL/NAACL workshop on text summarization.

[16] Dredze, M., Wallach, H. M., Puller, D., & Pereira, F. (2008). Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces* IUI '08 (pp. 199–206). New York, NY, USA: ACM. URL: `http://dx.doi.org/10.1145/1378773.1378800`. doi:10.1145/1378773.1378800.

[17] Filatova, E. (2004). A formal model for information selection in multi-sentence text extraction. In *In Proceedings of the International Conference on Computational Linguistics (COLING* (pp. 397–403).

[18] Fortes, G. L. F., de Lima Jos Valdeni, Loh, S., & de Oliveira, J. P. M. (2006). Using ontological modeling in a context-aware summarization system to adapt text for mobile devices. In P. P. Chen, & L. Y. Wong (Eds.), *Active Conceptual Modeling of Learning* (pp. 144–154). Springer volume 4512 of *Lecture Notes in Computer Science*.

[19] Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with wikipedia. *Artif. Intell.*, *194*, 130–150. URL: `http://dx.doi.org/10.1016/j.artint.2012.04.005`. doi:10.1016/j.artint.2012.04.005.

[20] Hamasaki, M., Matsuo, Y., Nishimura, T., & Takeda, H. (2009). Ontology extraction by collaborative tagging. In *WWW 2009*. ACM press.

[21] Hennig, L., Umbrath, W., & Wetzker, R. (2008). An ontology-based approach to text summarization. In *Web Intelligence/IAT Workshops* (pp. 291–294). IEEE.

[22] Hoffart, J., Yosef, M. A., Bordino, I., Frstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011). Robust disambiguation of named entities in text. In *EMNLP* (pp. 782–792). ACL.

[23] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, *46*, 604–632.

[24] Kogilavani, A., & Balasubramanie, B. (2009). Ontology enhanced clustering based summarization of medical documents. *International Journal of Recent Trends in Engineering*, *1*.

[25] Lau, R. Y. K., Song, D., Li, Y., Cheung, T. C. H., & Hao, J.-X. (2009). Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE Trans. on Knowl. and Data Eng.*, *21*, 800–813. doi:http://dx.doi.org/10.1109/TKDE.2008.137.

[26] Li, L., Wang, D., Shen, C., & Li, T. (2010). Ontology-enriched multi-document summarization in disaster management. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), *SIGIR* (pp. 819–820). ACM.

[27] Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (pp. 71–78).

[28] Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization* MMIES '08 (pp. 17–24). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: `http://dl.acm.org/citation.cfm?id=1613172.1613178`.

[29] Mittal, J. G. V., Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *In Proceedings of the ANLP/NAACL Workshop on Automatic Summarization* (pp. 40–48).

[30] Mohamed, A., & Rajasekaran, S. (2006). Improving query-based sum-

marization using document graphs. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on* (pp. 408–410). doi:10.1109/ISSPIT.2006.270835.

[31] Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *EMNLP* (pp. 763–772). ACL.

[32] Pang-Ning, T., Michael, S., & Vipin, K. (2005). Intoduction to data mining.

[33] Park, S., & Cha, B. (2008). Query-based multi-document summarization using non-negative semantic feature and nmf clustering. In *Networked Computing and Advanced Information Management, 2008. NCM '08. Fourth International Conference on* (pp. 609–614). volume 2. doi:10.1109/NCM.2008.246.

[34] Ping, C., & M., V. R. (2006). A query-based medical information summarization system using ontology knowledge. In *CBMS* (pp. 37–42). IEEE Computer Society.

[35] Pourvali, M., & Abadeh, M. S. (2012). Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base. *CoRR*, *abs/1203.3586*.

[36] Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 2004.

[37] Radev, D. R., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, *40*, 919 – 938.

[38] Rotem, N. (2011). Open text summarizer (ots). retrieved from http://libots.sourceforge.net/ in july 2011.

[39] Saravanan, M., & Ravindran, B. (2010). Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artif. Intell. Law*, *18*, 45–76. URL: http://dx.doi.org/10.1007/s10506-010-9087-7. doi:10.1007/s10506-010-9087-7.

[40] Steinberger, J., Kabadjov, M., Steinberger, R., Tanev, H., Turchi, M., & Zavarella, V. (2011). Jrc's participation at tac 2011: Guided and multilingual summarization tasks. In *TAC'11: Proceedings of the The 2011 Text Analysis Conference*.

[41] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* WWW '07 (pp. 697–706). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1242572.1242667. doi:10.1145/1242572.1242667.

[42] Takamura, H., & Okumura, M. (2009). Text summarization model based on the budgeted median problem. In *Proceeding of the 18th ACM conference on Information and knowledge management* (pp. 1589–1592).

[43] TexLexAn (2011). Texlexan: An open-source text summarizer. retrieved from http://texlexan.sourceforge.net/ in july 2011. URL: http://texlexan.sourceforge.net/.

[44] Thakkar, K., Dharaskar, R., & Chandak, M. (2010). Graph-based algorithms for text summarization. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on* (pp. 516 –519). doi:10.1109/ICETET.2010.104.

[45] Town, C. (2006). Ontological inference for image and video analysis. *Machine Vision and Applications*, *17*, 94–115. URL: http://dx.doi.org/10.1007/s00138-006-0017-3. 10.1007/s00138-006-0017-3.

[46] Wan, X., & Yang, J. (2006). Improved affinity graph based multi-document summarization. In *In Proceedings of HLT-NAACL, Companion Volume: Short Papers* (pp. 181–184).

[47] Wang, D., & Li, T. (2010). Document update summarization using incremental hierarchical clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 279–288).

[48] Wang, D., Zhu, S., Li, T., Chi, Y., & Gong, Y. (2011). Integrating document clustering and multidocument summarization. *ACM Trans. Knowl. Discov. Data*, *5*, 14:1–14:26. URL: http://doi.acm.org/10.1145/1993077.1993078. doi:http://doi.acm.org/10.1145/1993077.1993078.

[49] Wikipedia (2013). Wikipedia website. last access: 01/03/2013. URL: http://www.wikipedia.org.

[50] Wu, C.-W., & Liu, C.-L. (2003). Ontology-based text summarization for business news articles. In N. C. Debnath (Ed.), *Computers and Their Applications* (pp. 389–392). ISCA.

[51] Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., & Li, J. (2011). Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* SIGIR '11 (pp. 255–264). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2009916.2009954. doi:http://doi.acm.org/10.1145/2009916.2009954.

[52] Zhu, J., Wang, C., He, X., Bu, J., Chen, C., Shang, S., Qu, M., & Lu, G. (2009). Tag-oriented document summarization. In *Proceedings of the 18th international conference on World wide web* WWW '09 (pp. 1195–1196). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1526709.1526925. doi:http://doi.acm.org/10.1145/1526709.1526925.