



RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D3.2.1

Diversity-aware summarization

Editor:	Delia Rusu, JSI
Author(s):	Delia Rusu, JSI; Mitja Trampus, JSI; Andreas Thalhammer, UIBK
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	March 2013
Actual Delivery Date:	March 2013
Suggested Readers:	developers working on WP4 – Diversity Toolkit, developers creating case study prototypes in WP5
Version:	1.1
Keywords:	single document summarization, multi-document summarization

Disclaimer

This document contains material, which is the copyright of certain RENDER consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):

All RENDER consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):

All RENDER consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):

The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement. However, all RENDER consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):

The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the RENDER consortium as a whole, nor a certain party of the RENDER consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP3 Diversity representation and processing
Document Title:	D3.2.1 - Diversity-aware summarization
Editor (Name, Affiliation)	Delia Rusu, JSI
Work package Leader (Name, affiliation)	Andreas Thalhammer, UIBK

Copyright notice

© 2010-2013 Participants in project RENDER

Executive Summary

In this deliverable we describe three diversity-summarization perspectives, starting from fine grained entity summarization, continuing with single document summarization, and finally corpus summarization.

In the first part of this deliverable we present UIBK's survey of the state of the art in entity summarization, which has been published in the proceedings of the International Semantic Web Conference (ISWC) 2012. The paper has been annexed to the deliverable.

Next, we describe an approach to automatically generate single-document extractive summaries in a supervised manner, by selecting the most salient sentences from the original document. We investigate the usefulness of adding sentiment-bearing information as features for the summarizer, and conduct a comparative evaluation to assess the advantage of sentiment features. As training and testing data we make use of a standard dataset provided by an evaluation workshop.

Finally, we present a summarizer for generating extractive summaries from multiple input documents. We hereby investigate the feasibility of driving summarization with semantically represented inputs. The multi-document summarizer is integrated within the Google case study – DiversiNews, and the user can choose between obtaining summaries provided by this summarizer or the one developed by Google. The evaluation of the multi-document summarizer will be performed together with the evaluation of the Google case study and reported in D5.2.4 - Evaluation of the diversified news service.

Table of Contents

Executive Summary	3
Table of Contents	4
List of Figures.....	5
List of Tables.....	6
Abbreviations.....	7
Definitions	8
1 Introduction	9
2 Entity Summarization as an Enabling Technology for the Serendipity Effect.....	10
2.1 Entity Summarization and the Serendipity Effect.....	12
2.2 Survey on State-of-the-art Entity Summarization Approaches	13
3 Single-Document Summarization.....	14
3.1 Motivation.....	14
3.2 Algorithm	14
3.3 Evaluation.....	15
3.4 Integration	18
4 Multi-Document Summarization.....	19
4.1 Motivation.....	19
4.2 Algorithm	19
4.3 Integration	20
5 Conclusions and Future Work	21
References.....	22
Annex A Evaluating entity summarization using a game-based ground truth.....	23

List of Figures

Figure 1: Render examples for named entity representation.....	11
Figure 2: Screenshot of the online version of the New York Times.	11
Figure 3: Screenshot of Spiegel Online news service.	12
Figure 4: Mockup for presenting data-driven summaries.....	13
Figure 5: The summarization pipeline for a single document.....	14
Figure 6: ROUGE 2 evaluation results, using average recall, for the DUC 2007 dataset.....	16
Figure 7: ROUGE SU4 evaluation results, using average recall, for the DUC 2007 dataset.	17

List of Tables

Table 1. 10-fold cross validation on the DUC 2002 dataset, using only text analysis features.....	16
Table 2. 10-fold cross validation on the DUC 2002 dataset, using text analysis and sentiment features.	16
Table 3. ROUGE-2 and ROUGE-SU4 results for several systems for the update subtask of generating a summary of documents in cluster A. Our system id is 34, and the best system has the id 40.....	17

Abbreviations

SVM	Support Vector Machines
DUC	Document Understanding Conference
TAC	Text Analysis Conference
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

Definitions

DiversiNews the Google case study system. It is an interactive tool which allows users to browse and summarize news articles from different perspectives.

Enrycher the Diversity Mining Toolkit developed in WP2.

1 Introduction

In this deliverable we describe three diversity-aware summarization systems, each operating at different levels of granularity, and providing different summarization perspectives. We start with the finest level of granularity, and present several entity summarization approaches, which were published in a study on the state of the art of this field of research. Next, we describe a single document summarizer operating at document level. Finally, we present a multi-document summarizer generating corpus-level summaries.

Entity summarization is chosen to present topics and named entities in a way that show a variety of features without overloading the reader. The presented approaches to entity summarization emphasize features of the individual object rather than the category it belongs to. Annex A includes a survey on the state of the art of entity summarization approaches which was published in the proceedings of the International Semantic Web Conference (ISWC) 2012.

The single-document summarizer will be included within the Diversity Mining Services (Enrycher), and operates at document level. The summarization service completes the diversity mining pipeline, by adding an aggregated view of a document. The summarization algorithm automatically generates extractive summaries in a supervised manner, by selecting the most salient sentences from the original document. It uses a supervised machine learning approach based on support vector machines (SVM). The summarization task is seen as a binary classification one, where the classifier learns if a sentence belongs to the summary or not.

We investigate the usefulness of adding sentiment-bearing information as features for the summarizer, and conduct a comparative evaluation to assess the advantage of sentiment features. As training and testing data we use the Document Understanding Conference (DUC) dataset from 2007, comprising news articles from three news agencies.

The multi-document summarizer was implemented as an alternative to Google's summarizer. While Google's summarizer is based on probabilistic models, our multi-document summarization algorithm is based on text analysis features, as obtained from Enrycher. Our summarization algorithm generates extractive summaries from multiple documents at once. It integrates deeper into the data processing pipelines provided by RENDER to evaluate the possibilities of summarization based on semantic representation of input documents. Focused summarization is provided by the DiversiNews infrastructure (see Section 3) which acts as a pre-processing layer, filtering out the query-relevant information and feeding it to the summarizer.

The multi-document summarizer is integrated within the Google case study – DiversiNews, and the user can choose between obtaining summaries provided by this summarizer or the one developed by Google. The evaluation of the multi-document summarizer will be performed together with the evaluation of the Google case study and reported in D5.2.4 - Evaluation of the diversified news service.

This deliverable is structured as follows. Entity summarization viewed as technology enabler for the serendipity effect is presented in Section 2. Section 3 is dedicated to describing single document summarization. In section 3 we present the multi-document summarization algorithm. The final section of the deliverable is dedicated to concluding remarks.

2 Entity Summarization as an Enabling Technology for the Serendipity Effect

In contrast to the next sections which detail implementation aspects of approaches, this section covers motivation and background information of the new field of entity summarization. Hence, the outline and implementation of a specific entity summarization approach should be understood as a point for future work.

The field of Named Entity Recognition (NER) and linking to encyclopaedic sources is currently facing particular attention by news publishers, e.g. BBC news [1]. Thus, named entities are meant to be marked in the text such as demonstrated by the RENDER components Drupal Extension and Enrycher (see Figure 1). In many online newspapers such as the New York Times¹ (NYT), clicking on the named entities commonly navigates the user to a topic page. Note that this is coherent with the KDO ontology (cf. D3.1.1/D3.1.2) where we do not distinguish whether a topic reference marks a dmoz² topic or a named entity. Figure 2 shows an example for a topic page on the online version of NYT. The topic page includes a rather long textual description of the topic and further articles on the same topic. The latter is an essential part of news publishers' topic pages. However, as for the former, while the reader gets detailed background reading on the topic, the long text segments are very time consuming to read and therefore create an invisible barrier that could prevent users from consuming related articles or topics. As such, the textual summaries should be kept short and concise. For the automatic creation of text summaries the likeminded reader is referred to D3.2.1. An example for a more data-driven presentation of topics and named entities is provided by the German online newspaper Spiegel Online (SPON). Figure 3 shows this approach for the topic "Afghanistan". These "key fact" summaries of entities are currently focused on specific types of entities, in this case countries, and thus do not provide a great diversity. Despite the fact that it only counts for this specific types of entity the presented system also names the same properties for each country; this is rather monotonic as data-driven descriptions of countries include a huge variety of individual properties that could be presented shaped to the actual entity. All this leads us to the following questions:

- *Can we present topics and named entities in a way that show a variety of features without overloading the reader?*
- *Can this be done from a point of view that emphasizes on the features of the individual object rather than the category it belongs to?*

These questions lead us to the field of automatic entity summarization. Although quite new, the area is evolving fast and includes prominent examples such as the summaries of the Google Knowledge Graph. In the following sections we will firstly discuss on what entity summarization is and how it can help to feature the serendipity effect and, after that, we will investigate on the state of the art of entity summarization.

¹ <http://www.nytimes.com/>

² <http://www.dmoz.org/>

At least 2 tornadoes hit Dallas, Fort Worth

Submitted by [redacted]

DALLAS (USA Today) -- At least two tornadoes violently spun through the Fort Worth-Dallas-Fort Worth area, collapsing roofs, ripping down power lines and tossing trailers around like toys, authorities said Tuesday. The National Weather Service said "considerable damage" had been reported near Cleburne, south of Fort Worth, and Lancaster, south of Dallas. Local television footage showed overturned and smashed semi-trailers on the ground in the southern portion of Dallas County. "Obviously we're going to have a lot of assessments to make when this is done," Dallas County spokeswoman Maria Arita said. Flights heading to Fort Worth-Dallas-Fort Worth International Airport were being delayed almost three hours. AccuWeather reported significant damage at Six Flags Amusement Park in Arlington. Dallas Police spokeswoman Sherri Jeffrey said twisters also caused damage inside the city limits. The storms spewed hail, some as large as baseballs, the weather service said. John Nielsen-Gammon, Texas' state climatologist, says the tornadoes occurred when a high-level trough of cold air collided with surface warm air that had been floating over Texas from the Gulf of Mexico for days. Forecasters predicted severe weather for that area

Slovenia's dramatic win over Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near domination from the likes of Brazil, Italy and Germany, international soccer is entering the era of the Cinderella. It may not happen this time around, but given the increasing flow of talent, training and information across borders, it's almost certain that a small upstart nation blessed with good athletes and better luck will make a legitimate run at the world's most coveted trophy. Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Valter Birsa Wednesday. Wednesday near Paris, Ireland nearly pulled off the greatest win in its soccer history. As retired French star Zinedine Zidane watched in the crowd, the Irish exerted intense pressure on the 2006 finalists, missing four point-blank chances, one from two yards off the left foot of John O'Shea, the Manchester United defender. Striker Robbie Keane also missed from close range in the 73rd and 90th minutes. The misses would prove crucial in extra time, as French star Thierry Henry controlled a free kick from near midfield, tapping it twice with his hand before flicking a cross to the goal line that William Gallas headed into the net. Despite intense protests from Irish goalkeeper Shay Given, Swedish referee Martin Hansson allowed the score, giving France its one-goal edge. Russia wasn't so fortunate, losing 1-0 to tiny

Figure 1: Render examples for named entity representation.

The screenshot shows the New York Times website interface. At the top, there is a navigation bar with 'Times Topics' highlighted in a red box (1). Below the navigation bar, the main content area features a large article about Robert Menendez, with a photo and a red box (2) around the text. To the right of the article, there are sections for 'Headlines Around the Web', 'WHAT'S POPULAR NOW', and 'MOST POPULAR'. At the bottom of the article, there is an 'Articles' section with a red box (3) around it, listing related news items.

Figure 2: Screenshot of the online version of the New York Times: (1) shows that named entities are treated as topics, (2) provides a description of the topic, and (3) lists further articles with the same topic.

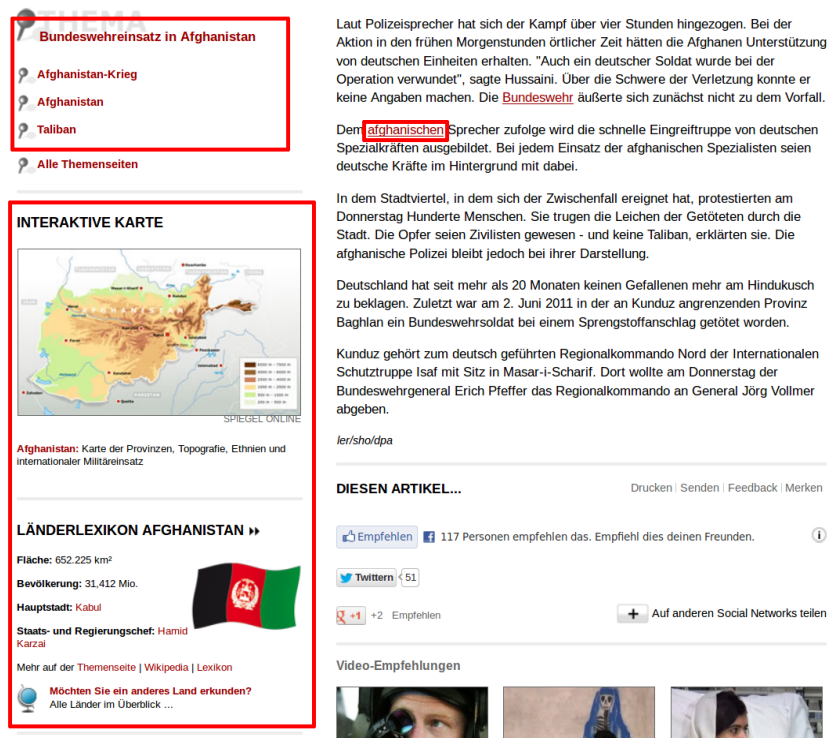


Figure 3: Screenshot of Spiegel Online news service. Note the background information with “key facts” on topics (in this case Afghanistan) as a sidebar.

2.1 Entity Summarization and the Serendipity Effect

The new field of entity summarization covers the aspect of lossy compression of information that is available in a graph structured way. Keeping the context of RDF and linked data in mind, a loose definition for the task could be stated as follows:

“Entity summarization is the task of producing a summary that conveys important facts about the entity while reducing the number of shown facts significantly” [3]

It has to be empathized that, although it might still be a method of summarization, using the category or type of an entity to represent all entities of this type with a fixed set of properties does certainly not suit the aim. For example, the feature “spouse” is certainly more important to be presented for “Andre Agassi” or “Nicolas Sarkozy” than for “Michael Schumacher” where “brother” might be more important. Thus, the summaries have to be focused on the specific and interesting features of the entity. Looking back to the definition of entity summarization, it is left to the reader to judge whether “interesting” and “important” can be used exchangeable in this context.

Serendipity as it is defined in Wikipedia “... means a ‘happy accident’ or ‘pleasant surprise’; specifically, the accident of finding something good or useful while not specifically searching for it” [4]. Entity summarization can feature this “happy accident” in multiple ways:

- The user is not overloaded with information while, at the same time, is attracted by the short and concise presentation.
- The change of properties for each and every individual prevents the user to get used to as well as bored by a fixed schematic pattern for the same type.
- The selection of important facts draws the user’s attention to the objects.

We believe that this technology could be an important factor in the process of presenting information in a diversified way. In the following we walk through an example that shows on how a summary of an entity can lead to the discovery of new and interesting aspects:

Figure 4 shows a mockup example which presents an example news Web page containing text mentioning the name “Bob Menendez”. After the NER step, the text fragment is linked to the DBpedia entity of Bob Menendez. In current DBpedia, there are about 376 facts contained about “Bob Menendez”. Presenting this information all at once would not make sense as it would overload the user with a myriad of uninteresting facts. Therefore, we made use of entity summarization in order to select 8 interesting facts that could be stated about this entity. For simplicity, these facts are presented in a table-like fashion contained in a box that opens itself when clicking on the entity as it occurs in the text. The box also contains a sentiment-oriented view showing a selection of other articles in which the entity occurs. The next question is to which destinations the shown property or object values should be linked. For the objects it makes sense to take the user to the object’s summary in the same window. This, again, presents a summary of the entity and shows the articles containing the object. In this case this would be summaries of New Jersey, John Kerry, etc. The click on one of the predicates (e.g. is dbpprop:senators of) could present other topics that have the same property-value pair (according to the example this would present the topic “Frank Lautenberg”, the second senator of New Jersey). To make clear, that property links link to other pages than object links, one could use a on-mouse-over effect with borders in order to make clear what could be expected when clicking on one of the links. As a final remark, it is important to note that for a summary which features serendipity aspects, literal values such as the date of birth (e.g. “1973-02-04”^^xs:date) cannot be linked to meaningful background information. Thus, we only present objects which are resources themselves.



Figure 4: Mockup for presenting data-driven summaries.

2.2 Survey on State-of-the-art Entity Summarization Approaches

As a complement to the ideas and descriptions in the previous sub-section, we would like to include the up-to-date state of the art of entity summarization approaches. For this, we refer to Section 2 of the RENDER publication [2] (see Annex A).

3 Single-Document Summarization

In this section we describe the single-document summarization algorithm that automatically generates extractive summaries. We take a classification approach to summarization, and train our summarization model using human-generated summaries provided by summarization evaluation workshops. Figure 5 shows the summarization pipeline for a single document. From each document we extract what we refer to as *text analysis features* and *sentiment analysis features*. Based on these sets of features, we train two classifiers: the **Text** classifier uses only text analysis features, while the **Sentiment** classifier uses both text analysis and sentiment analysis features. Using this approach, we can fully take advantage of the Enrycher extracted annotations. Both classifiers are Support Vector Machines (SVM) classifiers which were shown to provide good results in text classification [5]. The disadvantage of this approach is that it requires labelled data for training, which is sometimes time-consuming and expensive to obtain. There are, however, several summarization evaluation workshops (Document Understanding Conference³, Text Analysis Conference⁴) which provide labelled data, mainly from the news domain.

Section 3.2 describes in more detail the feature set as well as the classification algorithm.

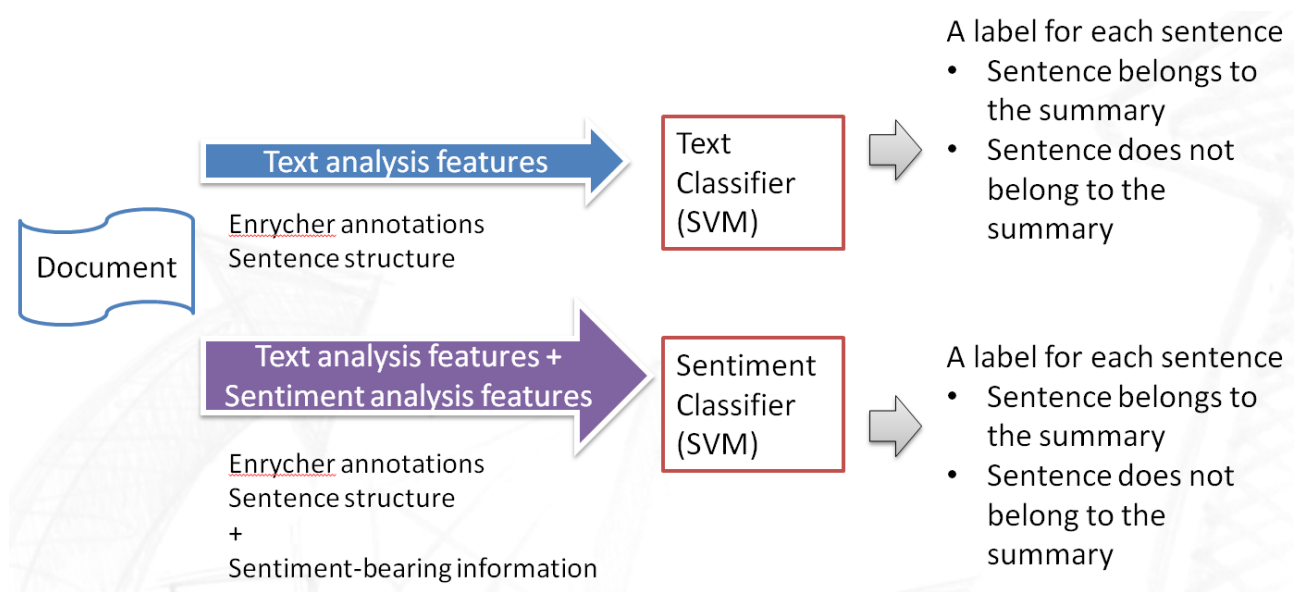


Figure 5: The summarization pipeline for a single document.

3.1 Motivation

Our goal is to compare the results obtained when using our summarization algorithm on two groups of features: a group containing no sentiment-bearing information, and a second one containing sentiment information. Our hypothesis is that by including sentiment-features we can improve the summarization results.

3.2 Algorithm

Our algorithm for automatically generating document summaries was inspired from Leskovec's work [7] and previously published in [8]. In this deliverable we present an extension of the algorithm in terms of the feature set which includes sentiment-bearing information. Given a document, we extract both text and sentiment analysis features from each of the text sentences, and train two linear SVM classifiers to identify which sentence should belong to the summary. Both text and sentiment analysis features are provided by Enrycher, the Diversity Mining Toolkit.

³ DUC (Document Understanding Conference): <http://duc.nist.gov/>, last checked on February 22, 2013.

⁴ TAC (Text Analysis Conference): <http://www.nist.gov/tac/>, last checked on February 22, 2013.

Text analysis features

We extract both features from the sentence, as well as features from the subject-predicate-object assertions extracted from each sentence.

Regarding *sentence-level features*, we consider:

1. named entities appearing in the sentence, broken down by entity type (locations, organizations, persons, dates, percentages, money).
2. Annotation semantics, as entities are linked to external linked open data sources such as DBpedia or OpenCyc
3. Parts of speech occurring in the sentence
4. Sentence similarity with a centroid (in this case we represented each sentence using a bag-of-words model, and determined the centroid of the document based on this bag-of-words representation of each sentence)

Regarding *assertion-level features*, we represent the extracted assertions as a graph, where the graph nodes are the subject and object assertions, and the edge between these nodes is represented by the verb. We can therefore determine, for each node its:

1. Page rank,
2. hub and authority weights,
3. length of chains starting from that node,
4. size of the connected component that node belongs to

Sentiment analysis features

We identify three sentiment analysis features for each sentence, namely:

1. if the sentence has attached sentiment information
2. the polarity of the sentence: positive, negative or neutral
3. the sentiment score, as given by the Sentiment Analysis Tool.

3.3 Evaluation

For summarization evaluation, we used the Document Understanding Conference (DUC) 2007 dataset. The DUC 2007 update summarization task provided a dataset consisting of 10 topics (A-J), each divided in 3 clusters (A-C), each cluster with 7-10 articles. We focused on the first part of the task – *producing a summary of documents in cluster A* – 100-words in length, without taking into consideration the topic information.

In order to obtain the 100-word summary, we first retrieved all sentences having triplets belonging to instances with the class attribute value equal to +1, and ordered them in an increasing manner, based on the value returned by the SVM classifier. Out of these sentences, we considered the top 15%, and used them to generate a summary. That is because most sentences that were manually labelled as belonging to the summary were among the first 15% top sentences.

The training data comprised 718 DUC 2002 documents, where the summary for each document was provided by human annotators for 147 of these documents.

Table 1. 10-fold cross validation on the DUC 2002 dataset, using only text analysis features.

DUC 2002	Test Set – using only text analysis features		
Documents	Precision	Recall	F1 measure
147	29.47%	73.03%	41.95%

Table 1 shows 10-fold cross validation results on the DUC 2002 dataset, using only text analysis features, while Table 2 shows results on the same dataset, this time using both text analysis and sentiment features. The results are similar to the ones reported in [7], on the same dataset. A slight improvement can be observed when combining both sentiment and text analysis features (see Table 2). However, most of the articles were annotated by the sentiment analysis tool as having neutral polarity. It was similar in the case of the DUC 2007 data. This is due to the fact that both DUC datasets contain news articles, which have a neutral way of reporting events.

Table 2. 10-fold cross validation on the DUC 2002 dataset, using text analysis and sentiment features.

DUC 2002	Test Set – using text analysis and sentiment features		
Documents	Precision	Recall	F1 measure
147	35.03%	71.60%	47.04%

We compared the performance of our system against various other systems, using ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [6], an automatic summarization evaluation package, which is available upon request⁵. ROUGE is frequently used in the DUC and TAC (Text Analysis Conference) evaluation series. We present results using the ROUGE 2 and ROUGE SU4 metrics, which were some the ones used in the DUC 2007 evaluation:

- **ROUGE-N:** N-gram based co-occurrence statistics;
- **ROUGE-SU:** Skip-bigram plus unigram-based co-occurrence statistics.

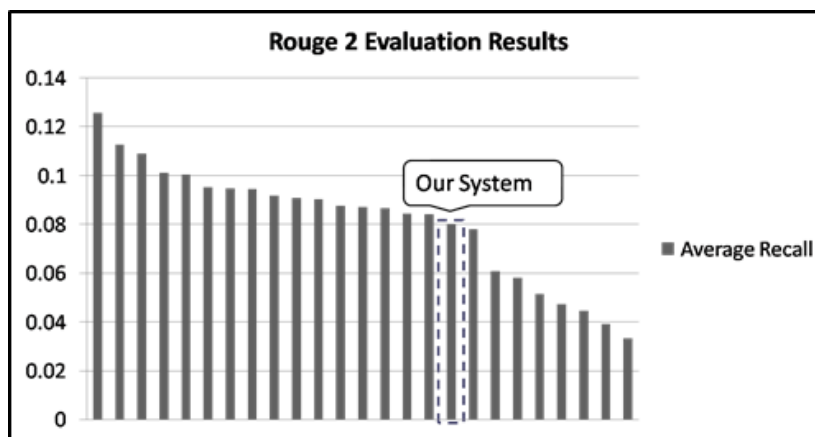


Figure 6: ROUGE 2 evaluation results, using average recall, for the DUC 2007 dataset.

On the DUC 2007 update task, our system was ranked 17 out of 25, based on the ROUGE-2 evaluation method, and 18 out of 25 based on the ROUGE-SU4 evaluation method (there were 25 systems participating in the 2007 update task). Changing the features to include sentiment analysis ones did not

⁵ ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*): <http://www.berouge.com/Pages/default.aspx>, last checked on February 22, 2013

modify the ranking. This is mainly due to the fact that our sentiment analysis system annotated many news articles as being neutral. This is not unexpected, as news articles are written in a more neutral manner, and include few sentiment-bearing words.

Figure 6 depicts the ROUGE 2 average recall obtained by our system, as well as the systems participating in the DUC 2007 update task. Figure 7 shows similar results, this time for the ROUGE SU4 evaluation metric. Note that these results were obtained using the evaluation script provided by the DUC 2007 conference organizers.

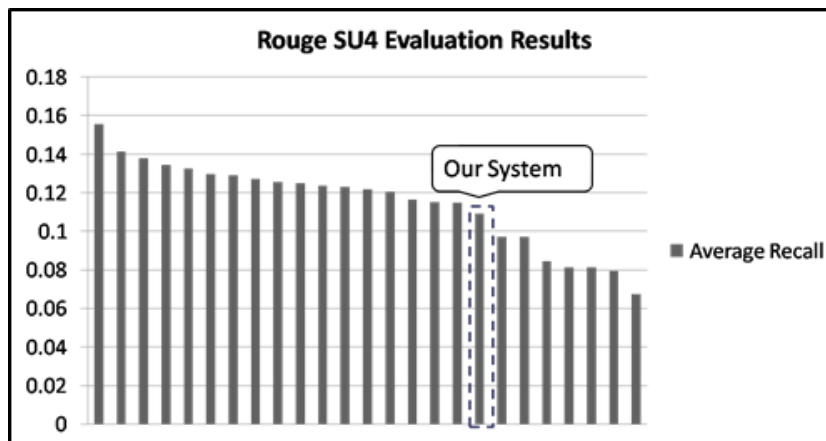


Figure 7: ROUGE SU4 evaluation results, using average recall, for the DUC 2007 dataset.

Table 3 shows ROUGE 2 and ROUGE SU4 results for several systems involved in the evaluation task, and also includes our system (ID 34). We show the three best performing systems (IDs 40, 51 and 46), as well as the worst (IDs 37 and 57). There are two baseline systems, with ids 35, corresponding to the system which returns all the leading sentences in the <TEXT> field of the most recent document and 58, corresponding to CLASSY04 [9], a system which obtained a high score in the DUC 2004 multi-document summarization task.

Table 3. ROUGE-2 and ROUGE-SU4 results for several systems for the update subtask of generating a summary of documents in cluster A. Our system id is 34, and the best system has the id 40.

System ID	ROUGE-2			ROUGE-SU4		
	Average Recall	Average Precision	Average F1	Average Recall	Average Precision	Average F1
40	0.126	0.125	0.125	0.156	0.155	0.155
51	0.100	0.103	0.102	0.130	0.134	0.132
46	0.092	0.089	0.090	0.126	0.122	0.124
38	0.087	0.087	0.087	0.122	0.121	0.121
34	0.080	0.079	0.080	0.109	0.108	0.109
37	0.051	0.055	0.053	0.081	0.087	0.084
57	0.033	0.034	0.034	0.067	0.068	0.068
35	0.046	0.052	0.048	0.082	0.095	0.088
58	0.086	0.082	0.084	0.123	0.118	0.120

Our system had an average performance on the DUC 2007 update dataset. However, note that we used very little training data, provided by the DUC 2002 summarization evaluation conference, which includes merely 147 annotated documents. Moreover, we did not optimize our system for the news domain, but rather kept it general, as in RENDER we are dealing with several kinds of data: news articles, tweets, blog entries, each with different characteristics.

3.4 Integration

The summarizer will be integrated within Enrycher, the Diversity Mining Toolkit developed in WP2.

4 Multi-Document Summarization

In this section, we present a multi-document summarization algorithm based on a semantic representation of input documents, compatible with the diversity ontology developed in RENDER and integrated with the remainder of the project software.

4.1 Motivation

The algorithm is built around two key assumptions. First, in the multi-document summarization setting, the strongest signal for the importance of a piece of information is that piece being repeatedly reported by multiple sources. In other words, the ideal summary is the intersection of input documents in the (here under-defined) semantic space. Second, transforming the documents into semantic space reduces the number of ways in which information can be represented, making it easier to detect repeating statements in input documents despite their different wording. Our representation of choice are the subject-verb-object triplets made available through Enrycher.

4.2 Algorithm

The presented algorithm performs extractive summarization, meaning that it selects a subset of sentences and presents them as the summary in an unchanged form.

The **preprocessing** step is to obtain a semantic representation for each of the documents, as already indicated above. This is done by extracting all subject-verb-object triplets from each of the sentences using Enrycher. In addition, all words are aligned to WordNet using lemmatization and the “first sense” heuristic for all words.

In the **first stage**, each triplet is scored separately. The score of a triplet is determined based on its position in the document and exhibits an exponential shape:

$$score(T) = \max(0.4, 0.8^{pos(p(T))}) / \sum_{p'} sim(p(T), p')$$

where T is the triplet, $p(T)$ is the sentence containing the triplet, $pos(p(T))$ is the zero-based index of the sentence within the document and $sim()$ is the similarity function. The similarity function is defined as the Jaccard similarity coefficient for the sets of character 4-grams of the two sentences and ranges from 0 (no similarity) to 1 (identity). The nominator quantifies the intuition that especially in news reporting, the important facts tend to be given early on. The denominator compensates for sentences that not only have similar content but are almost completely identical. This tends to happen frequently with journalistic texts as the content is often partially copied from a press release or a news syndication network’s article.

In the **second stage**, the triplets are connected into a directed weighted graph based on their *information flow* similarity. The *information flow* between two triplets is defined as the sum of pairwise flows between their constituents (subject, verb and object). The flow between two constituents X and X' is defined as follows:

- If the concepts are identical ($X==X'$), the flow is 1 in both directions.
- Otherwise, if X is a hypernym of X' , the flow from X to X' is 0.7 and from X' to X is 0.25.
- Otherwise, the flow between X and X' is zero.

In the **final stage**, the most relevant triplets are selected greedily. The sentences containing them are promoted into the summary. To determine the relevance of a triplet, we first define its *information content*. This is initialized to

$$IC(T) = score(T) + \sum_{T'} w_{T \rightarrow T'} score(T')$$

where IC is the information content, score was defined above and $w_{T \rightarrow T'}$ is the information flow as defined in the previous paragraph. The following two steps are then repeated greedily until the desired length of the summary is reached:

- Promote the shortest sentence containing the triplet T with the highest IC(T) into summary. Ignore sentences that were originally part of quoted speech or that begin with a linking word (“however”, “also”, ...). If no suitable sentences are found, skip T.
- For every triplet T contained in this sentence, decrease the IC(T') for all remaining input triplets T' by a factor of $(1 - w_{T \rightarrow T'})$.

The fact that IC(T) is based on the score (which in turn is based on frequency) ensures that the summary contains relevant information. The second step of the greedy iteration above ensures that the information contained in the summary is not redundant. Intuitively, the performed decrease in IC(T') can be understood as “if T is told, then $w_{T \rightarrow T'}$ of T' is already told as well, so its information content decreases by that much.”

The order of the output sentences is determined based on the sentences' positions in their respective original files. In case of ties, higher-scoring sentences are placed first.

The constants used in the algorithm were determined experimentally with a grid search and informal evaluation. More rigorous evaluation of the algorithm's performance is planned in the forthcoming deliverables as an integral part of the diversified news service use case.

4.3 Integration

The summarizer has been exposed as a standalone web service; the interface is documented at <http://aidemo.ijs.si/multidoc/>.

In addition, it has been integrated into DiversiNews (<http://aidemo.ijs.si/diversinews/>), the diversified news service use case demo.

5 Conclusions and Future Work

In this deliverable we presented three diversity-summarization perspectives, starting from fine grained entity summarization, continuing with single document summarization, and finally corpus summarization.

Entity summarization was chosen for its ability to present topics and named entities in a way that show a variety of features without overloading the reader. Several such systems were presented in the published survey of state of the art approaches to entity summarization.

The document summarizer was evaluated using standard datasets made available by summarization workshops, while the corpus summarizer was integrated within DiversiNews, the Google case study tool, and will be evaluated during this case study evaluation.

Adding sentiment-bearing information to the set of features used for learning a single document summarizer slightly improved the classification results. However, the experiments were carried out on a small dataset. Additional evaluation of the corpus summarizer on a much larger dataset will better show if diversity information can improve the summarization task.

References

- [1] Matt Shearer: BBC News Lab: Linked data, <http://www.bbc.co.uk/blogs/blogbbcinternet/posts/BBC-News-Lab>
- [2] A. Thalhammer, M. Knuth, and H. Sack: *Evaluating entity summarization using a game-based ground truth*, in *The Semantic Web – ISWC 2012* (P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, eds.), *Lecture Notes in Computer Science*, vol. 7650, pp. 350–361, Springer Berlin Heidelberg (2012). ISBN 978-3-642-35173-0
- [3] <http://www.slideshare.net/thalhamm/evaluating-entity-summarization-using-a-gamebased-ground-truth>, slide 5
- [4] <http://en.wikipedia.org/wiki/Serendipity>, last checked on February 22, 2013
- [5] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Proceedings of the European Conference on Machine Learning, Springer, 1998.
- [6] C-Y. Lin. *ROUGE: a Package for Automatic Evaluation of Summaries*. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.
- [7] J. Leskovec, N. Milic-Frayling, M. Grobelnik. *Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts*. National Conference on Artificial Intelligence (AAAI), 2005.
- [8] D. Rusu, B. Fortuna, M. Grobelnik and D. Mladenic. *Semantic Graphs Derived From Triplets with Application in Document Summarization*. *Informatica Journal* 33 (2009), no. 3, pp. 357 -- 362.
- [9] J. M. Conroy, J. D. Schlesinger, *Left-Brain/Right-Brain Multi-Document Summarization*, Proceedings of the Document Understanding Conference (DUC), 2004.

Annex A Evaluating entity summarization using a game-based ground truth

A. Thalhammer, M. Knuth, and H. Sack: **Evaluating entity summarization using a game-based ground truth**, in *The Semantic Web – ISWC 2012* (P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, eds.), *Lecture Notes in Computer Science*, vol. 7650, pp. 350–361, Springer Berlin Heidelberg (2012). ISBN 978-3-642-35173-0

Evaluating Entity Summarization Using a Game-Based Ground Truth

Andreas Thalhammer¹, Magnus Knuth², and Harald Sack²

¹ University of Innsbruck, Technikerstr. 21a, A-6020 Innsbruck
andreas.thalhammer@sti2.at

² Hasso Plattner Institute Potsdam, Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam
{magnus.knuth,harald.sack}@hpi.uni-potsdam.de

Abstract. In recent years, strategies for Linked Data consumption have caught attention in Semantic Web research. For direct consumption by users, Linked Data mashups, interfaces, and visualizations have become a popular research area. Many approaches in this field aim to make Linked Data interaction more user friendly to improve its accessibility for non-technical users. A subtask for Linked Data interfaces is to present entities and their properties in a concise form. In general, these summaries take individual attributes and sometimes user contexts and preferences into account. But the objective evaluation of the quality of such summaries is an expensive task. In this paper we introduce a game-based approach aiming to establish a ground truth for the evaluation of entity summarization. We exemplify the applicability of the approach by evaluating two recent summarization approaches.

Keywords: entity summarization, property ranking, evaluation, linked data, games with a purpose.

1 Introduction

The main idea of the Semantic Web is to make implicit knowledge explicit and machine processable. However, machines that process knowledge are not a dead end. In fact, after processing the returned results are either consumed by another machine or by human users. In this paper, we focus on the latter: the consumption of machine processed data by human users. A lot of efforts in the Semantic Web currently focus on Linked Data interfaces and Linked Data visualization. As for the former, most interfaces have been developed by the Linked Data community and usually show all information (usually as property-value pairs) that is available for an entity (e. g. Pubby¹, Ontowiki², etc.) and leave it to the user to decide which of the information is important or of interest. In May 2012, Google³ introduced its “Knowledge Graph” (GKG), which produces summaries for Linked Data entities. While it is not the first approach to rank properties or

¹ Pubby – <http://www4.wiwiss.fu-berlin.de/pubby/>

² Ontowiki – <http://ontowiki.net/>

³ Google – <http://google.com/>

features of Linked Open Data according to their relevance [9,11,3] the uptake by industry certainly gives incentives for further investigation in this subject. This has to be considered in line with the fact that Google processed 87.8 billion queries in December 2009 [4] which makes roughly 2.8 billion queries per day. Keeping the huge number of daily searches in mind, it was an interesting move by Google to devote a big part of its result pages to the GKG summaries. Having an average of 192 facts attached to an entity [3], producing a concise summary that is shaped to an entity’s individual characteristics states an interesting research problem.

In this paper we will discuss current developments in Linked Data entity summarization and fact ranking as well as the need for a gold standard in form of a reference dataset which makes evaluation results comparable. We introduce a novel application of games with a purpose (GWAPs) that enables us to produce a gold standard for the evaluation of entity summarization. We demonstrate the applicability of the derived data by evaluating two different systems that utilize user data for producing summaries (one of which is GKG). In the course of our explanations we will emphasize on the complete and correct description of our test settings and stress that all data (that does not violate the privacy of our users) is made publicly available.


The remainder of this paper is structured as follows: Section 2 gives a description of the state-of-the-art in Linked Data entity summarization including the Google Knowledge Graph. In Section 3 the processed data sets, the quiz game and the evaluated systems are explained in detail, while Section 4 reports the achieved results. Section 5 concludes the paper with a brief summary and an outlook on future work.

2 Background

In recent years, four approaches to Linked Data entity summarization have emerged including the one adopted by GKG. In the following, we will discuss all of those approaches and - in addition - present methods used for evaluating text summarization.

Google has introduced the “Knowledge Graph” in May 2012 [8]. The main idea is to enrich search results with information about named entities. In case of ambiguous queries, such as “lion king” (currently a musical and a film are returned), Google lists also different possibilities. Two examples for GKG summaries are shown in Fig. 1. Google’s summaries are usually structured as follows: After presenting the name of the entity and an attached plot (usually taken from Wikipedia) next to a picture, up to five “main facts” are listed. These facts differ heavily between entities of different RDF types but also – to a certain extent – between entities of the same RDF type. After that, for certain RDF types like architects or movies, domain-specific attributes such as ‘*Structures*’ (architects) or ‘*Cast*’ (movies) are presented. For those, Google also defines a ranking e. g. from left to right for the ‘*Cast*’ lists. In addition, a range of related entities is displayed (Google introduces this list with ‘*People also search for*’). In their blog, Google





Charles Rennie Mackintosh



Charles Rennie Mackintosh was a Scottish architect, designer, watercolourist and artist. He was a designer in the Arts and Crafts movement and also the main representative of Art Nouveau in the United Kingdom. Wikipedia






Born: June 7, 1868, Glasgow
Died: December 10, 1928, London
Spouse: Margaret MacDonald (m. 1900)
Periods: Vienna Secession, Glasgow School

Structures

House for an Art Lover Willow Tearooms The Lighthouse Glasgow School of Art


People also search for

Margaret MacDonald Frank Lloyd Wright Josef Hoffmann Ludwig Mies van der Rohe William Morris

Feedback






Inglourious Basterds



Inglourious Basterds is a 2009 war film written and directed by Quentin Tarantino and starring Brad Pitt, Christoph Waltz and Mélanie Laurent. Wikipedia






Initial release date: May 20, 2009
Directors: Eli Roth, Quentin Tarantino
DVD release date: December 15, 2009
Screenplay: Quentin Tarantino
Awards: BAFTA Award for Best Actor in a Supporting Role, [More](#)

Cast

Quentin Tarantino Christoph Waltz Mélanie Laurent Diane Kruger Eli Roth
Hans Landa Shosanna Dreyfus Bridget von Hammersmark Sgt. Donny Donowitz

People also search for

Pulp Fiction Django Unchained Reservoir Dogs The Hurt Locker X-Men: First Class
1994 2012 1992 2008 2011

Feedback

(a) GKG: architect and designer Charles Rennie Mackintosh.

(b) GKG: movie titled “Inglourious Basterds”.

Fig. 1. Examples for GKG summaries (Source: <http://google.com/>)

developers describe summaries as one of “three main ways” to enhance search results with GKG information [8]. To automatically generate summaries, Google utilizes the data of their users, i. e. the queries, “[...] and study in aggregate what they’ve been asking Google about each item” [8]. We assume that these queries are in most cases “subject+predicate” queries, such as “lake garda depth”, or “subject+object” queries such as “the shining stanley kubrick”. In some cases also “subject+predicate+object” queries might make sense such as “jk rowling write harry potter”⁴. It is worth mentioning that using queries for determining the users’ average interest in facts also has some pitfalls. For example, the query “inglourious basterds quentin tarantino” (querying for a movie and one of its directors) not only boosts the ‘directed by’ property but also the ‘starring’ property for the movie’s relation to the person Quentin Tarantino. Unfortunately, this leads to the situation that the main actor (namely Brad Pitt) is not mentioned in the cast list while the director – who is known for taking minor roles in his movies and is doing so in this particular one – takes his position (see Fig. 1b).

Thalhammer et al. [9] explain how entity neighborhoods, derived by mining usage data, may help to discover relevant features of movie entities. The authors outline their idea that implicit or explicit feedback by users, provided by consuming or rating entities, may help to discover important semantic relationships between entities. Having established the neighborhood of an entity with

⁴ In fact, this query was suggested by Google Instant (<http://www.google.com/insidesearch/features/instant/about.html>).

methods adopted from item-based collaborative filtering [7], the frequency of a feature that is shared with its neighbors is likely to give an indication about the feature's importance for the entity. A TF-IDF-related weighting scheme is also adopted as some features are generally very common (e.g., provenance statements). Unfortunately, the authors do not provide an evaluation of their system and only provide some preliminary results. In the later sections, we will refer to this approach as UBEs (usage-based entity summarization).

The term of "entity summarization" was initially introduced by [3]. According to the authors, entity summarization is the task of identifying features that "not just represent the main themes of the original data, but rather, can best identify the underlying entity" [3]. We do not fully agree with this definition. Rather than selecting features that unambiguously identify an entity, we suggest to select features that are most interesting to present to a user. Of course, for many entities there is a significant overlap between the features that best identify an entity and features that are most interesting for the users. As a further contribution, the authors introduce the term "feature" as a property-value pair. The approach presented in [3] applies a "goal directed surfer" which is an adapted version of the random surfer model that is also used in the PageRank algorithm. The main idea is to combine informativeness and relatedness for the ranking of features. In the conclusion of [3], the authors state that "user-specific notion of informativeness [...] could be implemented by leveraging user profiles or feedback" in order to mitigate the problem of presenting summaries that help domain experts but are not as useful for average users. The presented approach does not utilize user or usage data in order to provide summaries. However, this information could be given implicitly by the frequency of in and out links.

Waitelonis and Sack explain how exploratory search can be realized by applying heuristics that suggest related entities [11]. Assume that a user is currently browsing the current US president's Linked Open Data description. Attached to the president's URI are properties such as `dbpedia-owl:residence`, `dbpprop:predecessor`, or `dbpedia-owl:party`. Obviously, these links are useful to show in the context of exploratory search. However, as there are more than 200 facts attached to the entity, the authors propose to filter out less important associations (i.e., provide summaries). To achieve this, they propose and evaluate eleven different heuristics and various selected combinations for ranking properties. These heuristics rely on patterns that are inherent to the graph, i.e. they do not consider usage or user data. The authors conduct a quantitative evaluation in order to find out which heuristic or combination performs best. The results show that some heuristics, such as the Wikilink and Backlink-based ones, provide high recall while Frequency and Same-RDF-type-based heuristics enable high precision. Trials with blending also showed that either precision or recall can be kept at a significant high level, but not both at the same time. Like in the approach of GKG, the predicate and the object are decoupled. While the introduced heuristics address the predicates, the data gathering for the evaluation focuses on the objects. As exemplified above, this leaves space for ambiguity. In the discussion, the authors argue that summaries should be considered in

a specific context (i. e., “what is the search task?”) and therefore quantitative measures might not provide the right means to evaluate property rankings.

[3] and [11] provide evaluations of their approaches. Both provide a quantitative as well as a qualitative evaluation. In the quantitative evaluation, both approaches base their evaluation on DBpedia⁵ excerpts comprised of 115 [11] and 149 [3] entities. These entities were given to a sufficient amount of users in order to establish a ground truth with human created summaries. To the best of our knowledge, the results of these efforts are not publicly available.

In the field of automatic text summarization, [1] discusses two possible ways for evaluating summaries: *human assessments* and *proximity to a gold standard*. Thus, in this area, not only a gold standard had to be created but also a way to measure closeness to such a reference. As entity summarization deals with structured data only, such proximity measures are not needed: to measure the similarity between a summary and a ground truth, we can make use of classic information retrieval methods such as precision/recall, Kendall’s τ and Spearman’s rank correlation coefficient.

3 Evaluating Entity Summarization

We attempt to create a ground truth for the task of entity summarization by utilizing data gained from a game with a purpose. We exemplify our approach in the domain of movies. Thus, our research hypotheses is as follows:

A game-based ground truth is suitable for evaluating the performance of summarization approaches in the movie domain.

Our assumption is that implemented approaches that provide summaries should perform significantly better than randomly generated summaries when measuring the correlation to the established ground truth. It is important to note that the relevance of facts for the task of summarization will be evaluated on the entity level. This means that the same properties, objects, or even property-value pairs are of different importance for different subjects. As a matter of fact, the importance of facts for an entity might vary given different contexts and summarization purposes. However, summarization also involves a certain level of pragmatics, i. e. trying to capture the common sense to address as many users as possible.

In the following we detail the restraints for the chosen domain, the design of the quiz game, the interpretation of the gained data, and the experimental setup for the evaluated systems.

3.1 Employed Dataset

In our evaluation, we focus on movie entities taken from Freebase⁶. This dataset contains a large amount of openly available data and – in contrast to DBpedia

⁵ DBpedia - <http://dbpedia.org/>

⁶ Freebase – <http://www.freebase.com/>

Listing 1. Property chain for defining a “hasActor” property.

```

1 <http://some-name.space/hasActor>
2 <http://www.w3.org/2002/07/owl#propertyChainAxiom> (
3 <http://rdf.freebase.com/ns/film.film.starring>
4 <http://rdf.freebase.com/ns/film.performance.actor> ).

```

and the Linked Movie Database (LinkedMDB)⁷ – very detailed and well curated information. Large parts of this dataset are also used by Google for its summaries [8]. For the evaluation, we have randomly selected 60 movies of the IMDb Top 250 movies⁸ and derived the Freebase identifiers by querying Freebase for the property `imdb_id`. With facts about 250 movies, it is difficult to achieve the mandatory number of game participants for sufficient coverage. Therefore, we have restricted the number of movies to 60. We have downloaded RDF descriptions of the movies and stored them in an OWLIM⁹ triple store with OWL2 RL¹⁰ reasoning enabled. This enables us to connect properties (such as actors) that are linked via reification (such as the ‘film-actor-role’ relationship) directly with property chain reasoning. An example for creating such an axiom is provided in Listing 1. We have created such direct links for actors, role names, achieved awards, budgets, and running times. As a matter of fact, not all properties are useful to be questioned in a game. Therefore, we make use of a white list. The list of selected movies, the used property chain rules as well as the property white list are available online (cf. Sec. 4.3).

3.2 *WhoKnows?Movies!* – Concept and Realization

We developed *WhoKnows?Movies!* [10], an online quiz game in the style of ‘*Who Wants to Be a Millionaire?*’, to obtain a ground truth for the relevance of facts. The principle of the game is to present multiple choice questions to the player that have been generated out of the respective facts about a number of entities. In this case we limited the dataset as described in Sec. 3.1. The players can score points by answering the question correctly within a limited period of time and lose points and lives when giving no or wrong answers.

As an example, Fig. 2 shows the question ‘*John Travolta is the actor of ...?*’ with the expected answer ‘*Pulp Fiction*’, which originates from the triple

```
fb:en.pulp_fiction test:hasActor fb:en.john_travolta .
```

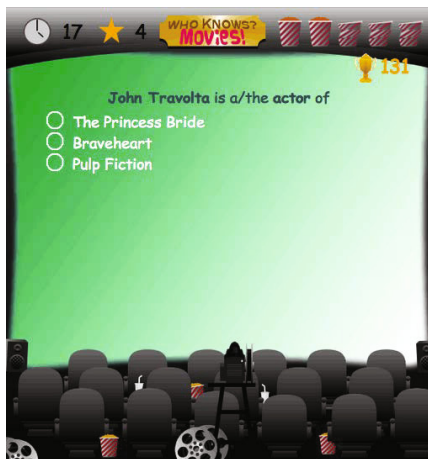
and is composed by turning the triple’s order upside down: ‘*Object is the property of: subject1, subject2, subject3...*’. The remaining options are selected from entities that apply the same property at least once, but are not linked to the object of the question. In this way we assure that only wrong answers are presented as alternative choices. There are two variants of questions: *One-To-One*

⁷ LinkedMDB – <http://www.linkedmdb.org/>

⁸ IMDb Top 250 – <http://www.imdb.com/chart/top>

⁹ OWLIM – <http://www.ontotext.com/owlim>

¹⁰ OWL2 RL – http://www.w3.org/TR/owl2-profiles/#OWL_2_RL



Subject	Property	Object
Pulp Fiction	actor actor actor	John Travolta Uma Thurman ...
Braveheart	actor actor actor	Mel Gibson Sophie Marceau ...
The Princess Bride	actor actor actor	Robin Wright Annie Dyson ...

Fig. 2. Screenshot and triples used to generate a One-To-One question

where exactly one answer is correct and *One-To-N* where one or more answers are correct.

When the player answers a question correctly he scores points and steps one level up, while incorrect answer will be penalized by losing points and one life. The earned score depends on the correctness of the answer and the time needed for giving the answer. With growing level the number of options raises, so correct answers are getting harder to guess. It has to be noted that the probability for a fact to appear in a question with many or few choices is equal for all facts. This ensures that the result is not skewed, for example by putting some facts in questions with two choices only. When submitting an answer, the user receives immediate feedback about the correctness of his answer in the result panel, where all choices are shown once again and the expected answer is highlighted. Given answers will be logged for later traceability and the triple's statistics are updated accordingly. The game finishes when the player lost all of his five lives.

Applying the white list described in Sec. 3.1, 2,829 distinct triples were produced in total. For each triple a set of wrong answers is preprocessed and stored into a database. When generating a question for a specific triple, a number of false subjects is randomly selected from this set.

3.3 What Are *Interesting Facts*?

The answer patterns of quiz games can tell a lot about what is generally interesting about an entity and what is not. One of the questions in the quiz game of Sec. 3.2 is 'What is the prequel of *Star Wars Episode VI*?' with one of the answer options being '*Star Wars Episode V*'. Of course, most of the players were right on this question. On the other hand fewer players were right on the question whether '*Hannibal rising*' is a prequel of '*The silence of the lambs*'. The idea of

a good general¹¹ summary is to show facts that are common sense but not too common. This is related to Luhn’s ideas about “significance” of words and sentences for the task of automatically creating literature abstracts [6]. Transferring the idea about “resolving power of words” to the answer patterns of the quiz game, we can state that neither the most known nor the most unknown facts are relevant for a good summary, it is the part between those two. Unfortunately, we have not been able to accumulate enough data to provide a good estimation for fine grained upper and lower cut-off levels. Therefore, in Sec. 4 we measure the relevance correlation with a pure top-down ranking.

In addition, there might be questions, where not knowing the right answer for a given fact does not necessarily mean that this fact does not have any importance. For our movie quiz game, participants are also asked for actors of a given movie. First of all, Freebase data does not distinguish between main actors and supporting actors. Thus, the property actor might not be in general considered as an important property, because most people do not know many of the supporting actors. Furthermore, an actor might play a very important role in a movie, but the game players do not know his name, because they only remember the face of the actor from the movie. The same holds for music played in the movie, where the participants might not know the title but are familiar with the tune. Thus, for future use, also the use of multimedia data should be considered to support the text-based questions of the quiz game.

3.4 Evaluated Systems

We exemplify the introduced evaluation approach to the summaries produced by GKG [8] and UBES [9]. For both approaches the additional background data stems from user behavior or actions. In addition, the rationale of both systems is to present useful information to the end users in a concise way. These similarities guarantee a comparison on a fairly equal level. In this section, we will detail the experimental setup and the data acquisition¹².

Usage-Based Entity Summarization (UBES)

In addition to Freebase, the UBES system utilizes the usage data of the HetRec2011 MovieLens2k dataset [2]. With a simple heuristic based on IMDb identifiers, more than 10,000 out of 10,197 HetRec2011 movies have been matched to Freebase identifiers (cf. [9] for more information). Based on the rating data provided by HetRec2011, the 20 nearest neighbors for each of the 60 selected movies were derived with the help of the Apache Mahout¹³ library. It has to be noted that the actual numerical ratings were not used due to utilization of the log-likelihood similarity score [5]. This similarity measure only uses binary information (i. e., rated and not rated). With two SPARQL queries per movie, the

¹¹ As opposed to contextualized and/or personalized.

¹² The final results of the UBES and GKG summaries, both using Freebase URIs, can be found in the dataset, cf. Sec. 4.3.

¹³ Apache Mahout – <http://mahout.apache.org/>

number of shared features was estimated once in combination with the neighbors and once considering the whole dataset. These numbers enable to apply the TF-IDF-related weighting for each property as it is described in [9]. Finally, the output has been filtered with the white list described in Sec. 3.1 in order to fit with the properties of the game and GKG.

Google’s Knowledge Graph (GKG) Summaries

The 60 movie summaries by Google have been processed in a semi-automatic way to fit with the Freebase URIs. The first step was to retrieve the summaries of all 60 movies and storing the according HTML files. While the Freebase URIs for properties such as “Director” had to be entered manually, most objects could be linked to Freebase automatically. For this, we made use of the GKG-Freebase link¹⁴. The ranking of the five main facts is to be interpreted in a top-down order while Google’s ordering of ‘Cast’ members follows a left to right orientation.

4 Results

At present, our quiz has been played 690 times by 217 players, while some players have played more frequently and the majority of 135 players has played only once. All 2,829 triples have been played at least once, 2,314 triples at least three times. In total 8,308 questions have been replied of which 4,716 have been answered correctly. The current results have to be regarded with care, since the absence of multiple opinions about a portion of the facts increases the probability for outliers. The random summaries were generated in accordance to the white list (cf. Sec. 3.1). In order to gain real randomness, we averaged the scores of 100 randomly generated summaries.

The ratio of correctly answered questions varies depending on the property that has been used in the question. As shown in table 1, to determine a movie according to its *prequel*, *film series*, or *sequel* is rather obvious, whereas a *film festival* or *film casting director* does not give a clear idea of the movie in question.

4.1 Evaluation of Property Ranking

To evaluate the ranking of properties for a single movie, we have determined the ranking of properties according to the *correct answer ratio*. The GKG movie representation lists general facts in an ordered manner, whereas the cast of the movie is displayed separately. Accordingly, only the remaining 24 properties are used for this evaluation. Properties that do not occur in the systems’ results are jointly put in the bottom position. For benchmarking the ordering of both summaries, Kendall rank correlation coefficient is applied. For each movie τ is determined over the set of its properties. Table 2 shows the average, minimum, and maximum findings of τ . It can be seen, that both systems as well as random

¹⁴ <http://lists.w3.org/Archives/Public/semantic-web/2012Jun/0028.html>

Table 1. Overall Relevance Ranking for Movie Properties

Rank	Property	Correct	Rank	Property	Correct
1	prequel	95.39%	14	production company	56.10%
2	film series	95.16%	15	runtime	54.52%
3	sequel	85.33%	16	music	54.11%
4	parodied	76.47%	17	award	53.41%
5	adapted original	74.32%	18	actor	52.86%
6	subject	73.91%	19	story writer	51.18%
7	genre	65.14%	20	editor	50.00%
8	initial release date	65.14%	21	event	50.00%
9	director	63.51%	22	cinematographer	44.20%
10	rating	61.61%	23	budget	42.78%
11	writer	61.61%	24	film festival	42.27%
12	featured song	60.00%	25	film casting director	41.32%
13	featured filming location	60.00%			

Table 2. Performance for Movie Property Ranking for Selected Movies

	τ_{avg}	τ_{min}	τ_{max}
UBES	0.045	-0.505 (The Sixth Sense)	0.477 (Reservoir Dogs)
GKG	0.027	-0.417 (The Big Lebowski)	0.480 (Reservoir Dogs)
Random	0.031	-0.094 (American Beauty)	0.276 (Monsters Inc)

perform equal in average. In each system, for about half of the movies the correlation is negative which means that the orderings are partly reverse compared ordering in the derived dataset. In general, none of the two systems' rankings differs significantly from a random ranking. This might be due to the sparsity of the dataset where most of the facts have been played only three times or less. Another negative influence might come from the fact that we aggregate on objects as we rank properties only and do not consider full property-value pairs.

4.2 Evaluation of Feature Ranking

For this evaluation the relevance ranking of the movie cast is compared to the user generated ground truth. Table 3 presents the average, minimum, and maximum findings of τ for the ranking of actors for a distinct movie. The results for the actor ranking are fairly equal for both systems in the average case. The average τ value differs from random scores. We have estimated that the difference to the random ranking is significant ($p < 0.05$) for both systems. This result provides an indication that the relative importance of property-value pairs can be captured by the statistics established through the game. It has to be mentioned, that - in some cases - the UBES heuristic provides none or very few proposals due to the required 'Cast' overlap to neighboring movies.

Table 3. Performance for Actor Ranking for Selected Movies

	τ_{avg}	τ_{min}	τ_{max}
UBES	0.121	-0.405 (The Princess Bride)	0.602 (Indiana Jones and the last Crusade)
GKG	0.124	-0.479 (The Princess Bride)	0.744 (The Matrix)
Random	0.013	-0.069 (Fargo)	0.094 (Good Will Hunting)

4.3 Published Dataset

By publishing the data collected within the game¹⁵, we encourage other researchers to apply this information for their purposes. The dataset consists of two main parts: first the aggregated statistics, which comprises the selected RDF triples and the respective players' performance. And second an anonymized log about the completed games that allows replay of user sessions with complete questions and results. Updates of these files will be published on a regular basis.

5 Conclusion and Future Work

In this paper a crowd sourcing approach implemented as a game with a purpose is demonstrated to gather relevance information about facts within a knowledge base and to establish ground truth data for evaluating summarization. We found indications that such a dataset can fulfill this purpose. However, the established dataset in its current state is too sparse to make valid assumptions about the importance of single facts.

Future development of the *WhoKnows?Movies!* game will also include images to help players to identify persons related to a movie, or other composed information artifacts. We also consider scoring properties that were listed in combination with an incorrect object while the user did not vote for this answer possibility. This is due to the fact that the user probably could exclude this possibility as he knew the correct object(s). Further research directions are increasing the number of movies and exploiting further domains. As for the latter, we consider the domains of books, music, places, and people. In principle, any domain where general knowledge is widely spread can be targeted with the game.

Acknowledgements. The authors would like to thank Ontotext AD for providing OWLIM-SE 5.0. This research was partly funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257790 (RENDER project).

¹⁵ The dataset is available at <http://yovisto.com/labs/iswc2012/>

References

1. Amigó, E., Gonzalo, J., Peñas, A., Verdejo, F.: Qarla: a framework for the evaluation of text summarization systems. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, pp. 280–289. Association for Computational Linguistics, Stroudsburg (2005)
2. Cantador, I., Brusilovsky, P., Kuflik, T.: 2nd ws. on information heterogeneity and fusion in recommender systems (hetrec 2011). In: Proc. of 5th ACM Conf. on Recommender systems, RecSys 2011. ACM, New York (2011)
3. 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)
4. comScore. comscore reports global search market growth of 46 percent in 2009 (2010), http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009
5. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74 (1993)
6. Luhn, H.P.: The automatic creation of literature abstracts. *IBM J. Res. Dev.* 2(2), 159–165 (1958)
7. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proc. of the 10th Int. Conf. on World Wide Web, WWW 2001, pp. 285–295. ACM, New York (2001)
8. Singhal, A.: Introducing the knowledge graph: things, not strings (2012), <http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>
9. Thalhammer, A., Toma, I., Roa-Valverde, A.J., Fensel, D.: Leveraging usage data for linked data movie entity summarization. In: Proc. of the 2nd Int. Ws. on Usage Analysis and the Web of Data (USEWOD 2012) co-located with WWW 2012, Lyon, France, vol. abs/1204.2718 (2012)
10. Waitelonis, J., Ludwig, N., Knuth, M., Sack, H.: WhoKnows? – evaluating linked data heuristics with a quiz that cleans up DBpedia. *Int. Journal of Interactive Technology and Smart Education (ITSE)* 8(3), 236–248 (2011)
11. Waitelonis, J., Sack, H.: Towards exploratory video search using linked data. *Multimedia Tools and Applications* 59, 645–672 (2012), 10.1007/s11042-011-0733-1