Query Based Arabic Text Summarization

[1]Ibrahim Imam, [2]Nihal Nounou, [3]Alaa Hamouda, [4]Hebat Allah Abdul Khalek
[1,4]Dept. of CS, Arab Academy for Science, Tech. and Maritime Transport, Cairo, Egypt
[2]Horizons Software, Haroun Al Rashid Street, Heliopolis, Cairo, Egypt
[3]Dept. of Computer Engineering, Al-Azhar University, Nasr City, Cairo, Egypt

Abstract

With the problem of increased web resources and the huge amount of information available, the necessity of having automatic summarization systems appeared. Since summarization is needed the most in the process of searching for information on the web, where the user aims at a certain domain of interest according to his query, in this case domain-based summaries would serve the best. Despite the existence of plenty of research work in the domain-based summarization in English, there is lack of them in Arabic due to the shortage of existing knowledge bases. In this paper we introduce a query based, Arabic text, single document summarization using an existing Arabic language thesaurus and an extracted knowledge base. We use an Arabic corpus to extract domain knowledge represented by topic related concepts/keywords and the lexical relations among them. The user's query is expanded once by using the Arabic WordNet thesaurus and then by adding the domain specific knowledge base to the expansion. For the summarization dataset, Essex Arabic Summaries Corpus was used. It has many topic based articles with multiple human summaries. The performance appeared to be enhanced when using our extracted knowledge base than to just use the WordNet.

Keywords

Arabic Text Summarization, Knowledge-Based Summarization, Query Expansion, Ontology Extraction From Text, Arabic Wordnet

I. Introduction

Due to the increased access of data on the web, it became harder to understand a certain topic without doing an effort of reading long documents and going through a lot of web pages to determine the most relevant ones. There came the need for automatic systems that would save the user's time, such as document clustering software, automatic summarizer, data mining software, etc. A summary can be defined as a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. The main goal of a summary is to present the main ideas in a document in less space [1].

Many researches have been done in the field of automatic text summarization revealing different types of it. The most famous type is the extractive summarization, which takes as input a collection of text fragments (paragraphs or sentences) for one or more documents, and selects some subset into a summary. The fragments are then ranked, which allows meeting different summary lengths requirements. The extraction techniques vary from using some important features to rank the fragments, to using a user's query to help indicating the fragments' importance, and to using a knowledge base to indicate the fragments' relevance to a specified domain.

When an online search engine is used, the results are expected to be of a certain domain, specified by a given query. There appeared the need of query-based summaries. Researches showed that using a knowledge base in the extractive summarization, improved results, such as [2], they used knowledge from WordNet1 as well as from UMLS, a medical ontology, and this improved the performance. Some researchers preferred to manually build their knowledge, such as [3] the authors manually identified a list of medical cue phrases and terms from a corpus of medical news articles, the sentences are then ranked using important features plus adding the presence of domain specific phrases. And also [4] used encyclopedic knowledge in Wikipedia to expand user's query. Some other researchers preferred to use ontologies to represent their knowledge, some use existing ontologies such as, [5] and [6] the user's query is expanded with synonyms and semantically related concepts using online available medical ontologies. And some other researchers constructed their ontology manually, such as [7] where the authors manually construct an ontology for a small domain of news articles, using the category labels of the ontology tree to score paragraphs.

In this paper we propose an Arabic text query based, single document summarizer using knowledge base. We used the Arabic WordNet (AWN) to provide us with the words' lexical synonyms for the query expansion. We also developed our own knowledge base consisting of the domain ontology concepts and relations among them. The user's query is expanded using the knowledge base as well, the sentences are ranked according to their relevance to the original and the expanded query, and finally the highest ranked sentences are included in the summary according to the desired length. We used the ESSEX the Arabic summaries corpus for testing. It has many articles grouped by their topic, and each article has multiple generated human summaries. We focused on the "Art & Music" and the "Environment" topics. The corpus, from which the knowledge base is extracted, is collected from the World Wide Web.

II. Review

Before explaining how the system works, some main aspects should be presented.

A. Ontology

The ontology is an explicit, formal specification of a shared conceptualization of a domain of interest. It should be restricted to a given domain of interest and therefore model concepts and relations that are relevant to a particular task or application domain [8]. Ontologies provide a richer knowledge representation that improves machine interpretation of data [9].

Manual acquisition of ontologies is a tedious and cumbersome task. It requires an extended knowledge of a domain and in most cases the result could be incomplete or inaccurate. Manually built ontologies are expensive, error-prone, and biased towards their developer. Researchers try to overcome these disadvantages of manual building of ontologies by using semi-automatic or automatic methods for building the ontology. Automation of ontology construction not only reduces costs, but also results in an ontology that better matches its application. During the last decade, several ontology learning approaches and systems have been proposed. They try to build ontology by two ways. One way is developing tools that are used by knowledge engineers or domain experts to build the ontology like Protégé and Jena, they are called the ontology modeling tools. Another way is semi-automatic or automatic building of ontologies by learning it from different information sources [10]. In the next two sub-sections we will talk about the two methodologies for automatic or semi-automatic ontology building.

B. Ontology Learning From Text

Ontology learning refers to extracting ontological elements (conceptual knowledge) from input and building ontology from them. It aims at semi-automatically or automatically building ontologies from a given text corpus with a limited human exert. The ontology building can be from scratch (automatic), or by adapting an existing ontology in a semi-automatic fashion using several sources [10].

Text or unstructured data is the most difficult type to learn from. It needs more processing than the semi-structured or structured data. The systems which have been proposed for learning from free text often consist of the following four main processes, although they differ in the methodology of each process:

1. NLP

An ontology extractor from text must perform some NLP processes on the corpus to be able to extract knowledge from it. In a matter of fact some pre-processing should be applied on the texts before NLP is, such as removing abbreviations, numbers, words that don't belong to the ontology language, diacritics (تشكيل) in case of Arabic, etc. NLP processes include POS taggers, parsers (shallow or dependency), NER (Named Entity Recognizer), removing stop words, and stemming or lemmatizing.

2. Concept Extraction

Concept or keyword extraction can be described as the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document [11]. In [11], the existing methods for keyword extraction were divided into four categories, (a) simple statistics such as term frequency. (b) Linguistic analysis such as POS tagging, the analysis are mostly combined with statistical measures. (c) Machine learning where a set of training documents are provided to the system which has a range of human-chosen keywords, then the gained knowledge is applied to find keywords from new documents. (d) Mixed approaches combines the methods mentioned or use some heuristic knowledge such as the position, length, layout features of the words, etc. These approaches can be applied on both, single word and multi-word concept extraction.

3. Relation Extraction

Ontologies, besides having a list of concepts should define the relations among concepts. Several researchers have attempted to find taxonomic relations expressed explicitly in texts by matching certain patterns which is referred to as Hearst-patterns. Other researchers have used the internal structure of noun phrases to find taxonomic relations [12]. Some relations are determined through the dependency parsers, "is-a" or "part-of" relations.

4. Ontology or Hierarchy Building

Some of the ontology extraction tools have to have reference ontology to update it with the new concepts and relations it deducted. Other tools use association rules of concepts and relations to be able to remove redundancies in the produced ontology hierarchy. And finally some tools use FCA formal concept analysis, which is a principled way of deriving a concept hierarchy or formal ontology from a collection of objects and their properties [13].

There are some available tools that extract ontology from text, such as Text-To-Onto, and its successor text2Onto, OntoLearn, protégé plugin OntoLT, TERMINAE and some other done by researchers such as CRCTOL and Automatic construction of ontology from Arabic texts. It's worth mentioning that none of them supports the Arabic language except for the last one, the authors aim at building ontology for the whole Arabic language, which is not the case in this paper, it also requires an existing ontology to update it with the rules it extracted.

C. AWN (Arabic WordNet)

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. Though WordNet contains a sufficiently wide range of common words, it does not cover special domain vocabulary, since it is primarily designed to act as an underlying database for different applications [14] . The AWN follows the methodology of the EuroWordNet [15].

III. Proposed System Description

The system contains two main components, the first component is responsible for the knowledge construction, and the second one does the summarization. Fig. 1 shows the proposed system architecture.

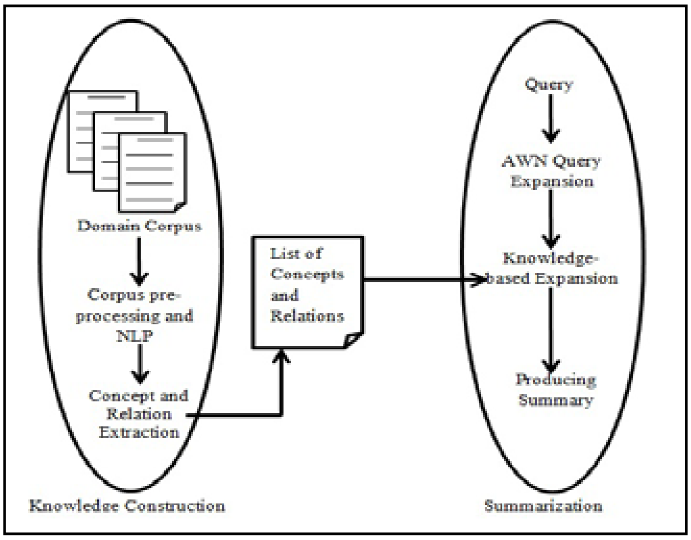
Domain Corpus
Corpus pre-processing and NLP
Concept and Relation Extraction
List of Concepts and Relations
Query
AWN Query Expansion
Knowledge-based Expansion
Producing Summary
Knowledge Construction
Summarization

Fig. 1 Proposed system architecture

A. Knowledge Construction

1. Corpus Pre-Processing and NLP

Different Arabic articles were collected from the internet, any certain domain articles are grouped together so they can be fed to the system, which gives the user the privilege of building knowledge to any desired domain. Some pre-processing is made to the corpus before knowledge extraction can be applied. Any none Arabic words or letters, numbers, diacritics (تشكيل), symbols or none letters such as brackets or quotations, extra spaces or empty lines are removed. Stop words aren't removed because they will be used in the concept and the relation extraction.

Stanford POS tagger is then used to determine the type of each word, i.e. noun, verb, etc. Stanford POS tagger is a Java implementation of the log-linear part-of-speech taggers described in [16]. Stanford POS tagger will be used in the concept and the relation extraction.

2. Concept Extraction

In our system we extracted the multi-word concepts, which are composed of words that co-occur together more often than can be expected by chance. We chose that because according to some studies such as [17], most domain-specific concepts are multiword terms. The small number of relevant single-word terms can either be found appearing frequently in the multiword terms or easily inferred based on the multiword terms. Also single word concepts may include general concepts as well as domain ones, and the relation extraction might not be so easy especially in the Arabic language, with the lack of dependency parsers and human intervention.

There are a lot of approaches to extract multi word keywords; a survey about that can be found in [18]. But just to mention them, some use statistical approaches then apply frames on the results to exclude the unwanted patterns such as [20] and [21], and some other approaches apply parsing or use POS tagger to choose certain patterns such as [17] and Text-To-Onto. All approaches then calculate the frequency of each pattern among the corpus, some of them stop at that step and some others extend the calculations to enhance performance such as the domain relevant measures (DRM), terminology identification measure (TIM) in [17], or the C-value/NC-value. In this paper we used the C-/NC value measures because in the previous research it showed better precision and recall values than to just depend on the frequencies [18]. And the C-/NC-value method is an efficient domain-independent multi-word term recognition method which combines linguistic and statistical knowledge [19].

The C-value/NC-value algorithm [22]:

Tagging the corpus.

Choose certain patterns for candidate concepts: Noun+Noun, (Adj|Noun)+Noun, (Adj|Noun)+ (Adj|Noun)* or with a prep among them. In the algorithm the maximum number of words to form a concept was not defined. But in Arabic language it is reasonable not to consider more than four words. It becomes more problematic to extract multi-word concepts of more than three words [21], and also in [20] they extracted only four.

A stop list which is a list of words that are not expected to occur as keywords in that domain, though they might exist more frequently. In the approach they studied the domain texts they had and produced a list, which will not be needed in query based summarization because the user's query eliminates the undesired concepts.

Calculate the total frequency of the candidate string in the corpus.

The frequency of the candidate string as part of other longer candidate terms.

The number of these longer candidate terms.

The length of candidate strings (in number of words).

Calculate the C-value.

Relation extraction: Techniques for taxonomic relation extraction vary from using a dependency parser to determine the relation among concepts to statistical approaches. Some tools use dependency parsers to determine if one concept is the subject of another, which indicates a relation, while other tools use shallow parsers or POS taggers and consider the following pattern (Noun1, Verb, Noun2) [17] depending on that verbs are hypothesized to indicate semantic relations between concepts. Some other tools consider the relations of "is-a" and "has-a" such as [20], and some tools use a mixture of that. All these tools also detect the frequency of concepts that appear together. Text2Onto [23], developed JAPE patterns for both shallow parsing and the identification of concepts and different types of relations. JAPE rules have to be developed by humans who are aware of the domain, and the rules are processed using GATE the NLP tool. Text2Onto supports only the English language for the easiness of JAPE rules creation.

Due to the lack of Arabic dependency parsers, in this paper we used the "is-a" relations such as: "هو، هي، هما، ...", "has-a" relations such as: "تتألف من، تتكون من، تنقسم إلى،..." and this list is prone to stretching if any new phrases were discovered. If any two concepts are mentioned together with a high frequency, it's considered a relation. Finally we adopted (Concept$_1$, Verb, Concept$_2$) approach to determine relations using the Stanford POS tagger.

In our approach we didn't build the ontology hierarchy, because it is out of the scope of the summarization and the hierarchy wouldn't add a value to the already built knowledge of concepts and relations.

B. Summarization

1. Corpus Collection and Pre-Processing

As mentioned before that several articles about different domains are collected from the internet, when the user enters his query, the corpus is searched for all the related articles to the query. The articles are then pre-processed in the same way we explained in the previous section.

2. AWN Query Expansion

The user's query is expanded by running it against the Arabic WordNet. AWN doesn't provide JAVA API like the EuroWordNet, so we had to use the AWN database and access the source code to retrieve all the synsets. If the user's query has any stop words, symbols, none letters or none words, they are removed before expansion. The user's approval for the expansion results is asked for, by removing any words that seem irrelevant.

3. Knowledge Base Query Expansion

The original query is expanded against the knowledge base of concepts and relations. All the related concepts to every word in the query and their relations are added. Also the expanded query is finalized by the user. The finalized and validated query is then used in the summarization.

4. Producing Summary

Every article is summarized as follows; each sentence in the corpus is given a score by comparing every word in it with the original and the expanded query. If the word exists in the original query it is given a score of "1" and if it is in the expanded query it is given the weight of "0.5". The score of the whole sentence is the summation of the words' scores. The list of sentences for each article is sorted in an ascending order. The user is asked for the preferred showed amount of summary, which is minimum 50% of the original document, but the user can choose less. The sentences with the highest scores are then displayed to the user in the same order they appeared in the original document.

IV. Evaluation and Results

Evaluation of summarization is a quite hard problem. Often, a lot of manual labor is required, for instance by having humans read generated summaries and grading the quality of the summaries with regards to different aspects such as information content and text clarity. Manual labor is time consuming and expensive. Summarization is also subjective. The conception of what constitutes a good summary varies a lot between individuals, and of course also depending on the purpose of the summary [30].

Some automatic text summarization tools use DUC (Document Understanding Conference) datasets to test their algorithms and some of them use human evaluation such as [24], while others use the abstract of an article as the human summary [6].

Due to lack of Arabic datasets or proper Arabic papers with abstracts, this approach [25], translated DUC datasets using Google translate. In our case we used The EASC, it is an Arabic natural language resources. It contains 153 Arabic articles and 765 human-generated extractive summaries of those articles. These summaries were generated using Mechanical Turk [26], which is a subsidiary of Amazon.com that provides a Web services system that uses people to perform tasks better handled by humans than computers [27]. This data also suited our domain knowledge base summarization; because the articles are divided into topics, i.e. art & music, education, environment, finance, health, politics, religion, science & technology, sports, and tourism. We used the environment domain in our testing, 105 articles were collected from the Internet, especially the Arabic Wikipedia, to be used in building the domain knowledge

For the evaluation we used the latest version of ROUGE 1.5, available in [28]. ROUGE is based on an n-gram co-occurrence between machine summaries and human summaries and is a widely accepted standard for evaluation of summarization tasks [29].

First we used the ROUGE to evaluate the human summaries, and then we compared our WordNet query expansion summaries with the human summaries once and then again after adding the knowledge-based query expansion.

Table 1: Human Summaries

	Avg_R	Avg_P	Avg_F
ROUGE 1	0.39027	0.33003	0.32908
ROUGE 2	0.21065	0.14641	0.16459
ROUGE-L	0.39027	0.33003	0.32908
ROUGE-SU4	0.24255	0.18428	0.17556

Table 2: Wordnet

	Avg_R	Avg_P	Avg_F
ROUGE 1	0.31218	0.34997	0.30805
ROUGE 2	0.10062	0.06388	0.07673
ROUGE-L	0.31218	0.34997	0.30805
ROUGE-SU4	0.11390	0.07102	0.08173

Table 3: After Adding Knowledge

	Avg_R	Avg_P	Avg_F
ROUGE 1	0.35919	0.42552	0.36569
ROUGE 2	0.19405	0.16389	0.16823
ROUGE-L	0.35919	0.42552	0.36569
ROUGE-SU4	0.20253	0.18750	0.16430

Table 1 shows the results of comparing multiple human summaries together. We used ROUGE 1, ROUGE 2, ROUGE L, and ROUGE SU4, and as shown that ROUGEs 1 and L produce better results. Table II shows the results of comparing the human summary with using only the WordNet in query expansion. Table III shows the results when adding the knowledge to the expansion. And it's shown that adding the knowledge base enhances the results.

V. Conlusion and Future Work

Our system is a query-based single document summarizer; it expands the user's preferable query using AWN. We also detected how adding a knowledge base to the query expansion would increase the performance of the system, which was proven by the evaluation results of ROUGE. To extract the knowledge, we used statistical and lingual measures on an existing domain corpus, getting a list of the domain's main concepts and the relations amongst them. We used EASC datasets for the evaluation, which is an Arabic articles grouped into certain domains, and each article has multiple human generated summaries.

The possible future work for this paper includes

- To build the ontology hierarchy, this might, using some complex association rules algorithms, detect new concepts relations that cannot be extracted explicitly from the corpus.
- Also if a similar approach could be found to compare the results with, or maybe translate some of the existing relative approaches data into Arabic, to perform better and wider range of comparison.
- Expand our work to multi document summarization.
- Add more extractive features to the summary.

References

[1] Dragomir R. Radev, Kathleen McKeown,"Introduction to the Special Issue on Summarization", Computational Linguistics – Summarization, Vol. 28, No. 4, pp. 399-408, 2002.

[2] Rakesh Verma, Ping Chen, Wei Lu,"Introduction to the Special Issue on Summarization", IEEE Transactions on Information Technology in Biomedicine, Vol 5, No. 4, pp. 261-270, 2007.

[3] Kamal Sarkar,"Using Domain Knowledge for Text Summarization in Medical Domain", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 200-205, 2009.

[4] Vivi Nastase,"Topic-Driven Multi-Document Summarization with Encyclopedic Knowledge and Spreading Activation", conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, 2008.

[5] A. A.Kogilavani, B. Dr.P.Balasubramanie,"Ontology Enhanced Clustering Based Summarization of Medical Documents", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 546-549, 2009.

[6] Ping Chen, Rakesh Verma,"A Query-based Medical Information Summarization System Using Ontology Knowledge", Computer-based Medical Systems (CBMS), 19th IEEE International Symposium, USA, pp. 37 – 42, 2006.

[7] Chia-Wei Wu, Chao-Lin Liu,"Ontology-based Text Summarization for Business News Articles", ISCA 18th International Conference on Computers and Their Applications, Honolulu, Hawaii, USA, pp. 389-392, 2003.

[8] Paul Buitelaar, Philipp Cimiano, Bernardo Magnini,"Ontology Learning from Text: Methods, Application and Evaluation,

IOS Press", 2003.

[9] Ivan Bedini, Benjamin Nguyen,“Automatic Ontology Generation: State of the Art”, Molecular Evolution, Vol. 44, No. 2, pp. 226-233, 1997.

[10] Maryam Hazman, Samhaa R El-Beltagy, Ahmed Rafea,“A Survey of Ontology Learning Approaches”, Vol. 22, No. 9, pp. 36-43, 2011.

[11] Elena Demidova, Iryna Oelze,“Automatic Keyword Extraction for Database Search”, Ph.D thesis, University of Hannover, 2009.

[12] Philipp Cimiano, Aleksander Pivk, Lars Schmidt-Thieme, Steffen Staab,“Learning Taxonomic Relations from Heterogeneous Evidence”, In: Ontology Learning from Text: Methods, Applications and Evaluation, pp. 59-73, IOS Press, 2005.

[13] Wikipedia, [Online] Available: http://www.en.wikipedia.org/wiki/Formal_concept_analysis, (10-01-2013).

[14] Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/WordNet, (10-01-2013).

[15] William BLACK, Sabri ELKATEB,“Introducing the Arabic WordNet Project”, Third International WordNet Conference (GWC-06), Korea, 2006.

[16] The Stanford Natural Language Processing Group, [Online] Available: http://www.nlp.stanford.edu/software/tagger.shtml, (14-01-2013).

[17] Xing Jiang, Ah-Hwee Tan,“Mining Ontological Knowledge from Domain-Specific Text Documents”, Data Mining, Fifth IEEE International Conference, Singapore, 2005.

[18] Euthymios Drymonas,“Exploring multi-word similarity measures for Information Retrieval applications: the T-SRM method”, Ph.D thesis, Technical University of Crete (TUC), Department of Electronics and Computer Engineering, 2006.

[19] Sophia Ananiadou, Hideki Mima,“An Application and Evaluation of the C/NC-value Approach for the Automatic term Recognition of Multi-Word units in Japanese”, International Journal of Terminology, Vol. 6, No. 2, pp. 175–194, 2000.

[20] Ahmed Cherif Mazari, Hassina Aliane, Zaia Alimazighi. “Automatic construction of ontology from Arabic texts”, ICWIT, Vol. 867, pp. 193-202, 2012.

[21] Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, “Automatic Extraction of Arabic Multiword Expressions”, the 7th Conference on Language Resources and Evaluation (LREC), 2010.

[22] Katerina Frantzi, Sophia Ananiadou, Hideki Mima, “Automatic recognition of multi-word terms: the C-value/NC-value method”, International Journal on Digital Libraries, Vol. 3, No. 2, pp. 115-130, 2000.

[23] Philipp Cimiano, Johanna Völker,“Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery”, 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Spain, pp. 227-238, 2005.

[24] Mahmoud O. EL-HAJ, Bassam H. HAMMO,“Evaluation of Query-Based Arabic Text Summarization System”, Natural Language Processing and Knowledge engineering International Conference, IEEE, Jordan, pp. 1-7, 2008.

[25] Mahmoud El-Haj, Udo Kruschwitz, Chris Fox, “Multi-Document Arabic Text Summarization”, Computer Science and Electronic Engineering Conference (CEEC), IEEE, UK, pp. 40 – 44, 2011.

[26] Summarisation Corpora, [Online] Available: http://privatewww.essex.ac.uk/~melhaj/easc.htm, (14-01-2013).

[27] PCMAG.com, http://www.pcmag.com/encyclopedia_term/0,1237,t=Mechanical+Turk&i=57289,00.asp, (14-01-2013).

[28] ROUGE, [Online] Available: http://www.berouge.com/Pages/Download ROUGE.aspx, (14-01-2013).

[29] Kavita Ganesan, ChengXiang Zhai, Jiawei Han,“Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions”, The 23rd International Conference on Computational Linguistics (COLING '10), China, 2010.

[30] Jonas Sjobergh,“Older versions of the ROUGEeval summarization evaluation system were easier to fool”, the International Journal of Information Processing and Management, Vol. 43, No. 6, pp. 1500-1505, 2007.