REPORT

on the paper

"OPINOSIS - A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions"

by

Kavita Ganesan , ChengXiang Zhai and Jiawei Han

Cosmina Croitoru and **Besnik Fetahu**

1 A (short) Critical Summary of the Paper

This paper, published in COLING 10 ([3]), introduces **Opinosis**, a lightweight framework for summarizing *highly redundant opinions* expressed in a multi-document input, to generate *concise informative abstractive summaries*.

This could be a valuable assistant in real life decision taking based on consumer surveys to assess a particular market or the attitude of its consumers. A concise textual summary of the redundant opinions expressed by the consumers comments complements and explains the quantitative information provided by the survey.

The paper hypothesis is that in order to summarize such corpora there is no need for prerequisites about domain knowledge or manual effort to apply deep NLP techniques: *the aggregate opinion expressed in the summary could be obtained only using the lexical links occurring in the sentences, ingeniously processed on a graph representation of the text.*

Evaluation results on a set of real data depict reasonable agreement with human summaries.

2 Background: Text Summarization

Text summarization was the first human approach to *Information overload*. It is therefore natural that *automatic summarization* to be one of important area of research of the Natural Language Processing (NLP) field.

Usually, the applications uses two kind of NLP, depending on the complexity of the tools employed.

Shallow NLP: mixing *simple syntactic features* (word order or location and similarity) with domain-specific interpretation; and

Deep NLP: sophisticated *syntactic, semantic and contextual processing*, e.g. named-entity recognition, relation detection, coreference resolution, syntactic alternations, word sense disambiguation, logic form transformation, logical inferences (abduction) and commonsense reasoning, temporal or spatial reasoning, etc.

At the address [7], there are a lot of open source NLP tools. For example, the *Part of Speech* annotation of any text could be obtained such that each word occurring in the text is annotated with a tag about *how the word is used* there. This is a shallow NLP use, and some examples are in the table bellow:

POS Tag	Meaning	Example
cc	coordinating conjunction	*and*
dt	determiner	*the*
nn	noun, singular	*table*
vb	verb, base form	*take*
jj	adjective	*red*
rb	adverb	*however, here, good*
in	preposition	*in, of, like*
to	TO	*to go, to him*

We are interested here in *text summarization* as a NLP tool to create of a shortened version of a text, maintaining its most important points. The goal is to obtain *coherent and correctly-developed* summaries, that is being *concise, readable,* and fairly *well-formed.*

Radev ([9]) identifies the following criteria to classify the various *types of summaries*:

Purpose - *Indicative, informative,* and *critical summaries*;

Form - *Extractive summaries* (salient paragraphs, sentences or phrases are extracted from the text);

- *Abstractive summaries* (find the central subject and produce a concise summary of this);

Dimensions - *Single-document summaries* vs. *multi-document summaries;*

Context - *Query-specific summaries* (usual in the search engines) vs. *query-independent summaries.*

Since the paper under review deals with *multi-document summarization*, some details are necessary. The goal in this case is to organize the information around the key aspects occurring in *all documents*, to represent a wider *diversity of views* on the topic. The methods developed for multi-document summarization use deep NLP, for example *centroid-based* and the use of *sentence utility* (MEAD [8, 9]) or based on *reformulation* ([6]) or based on *generation by selection and repair* ([1]).

MEAD (used in the Opinosis evaluation as a baseline) implements *extractive summarization:* selects a subset of highly relevant sentences from the cluster's overall set of sentences. It uses deep NLP and machine learning techniques, for example a *decision-tree* trained on a manually annotated corpus for CST relationships. The CST relationships mean *Cross-document Structure Theory* relationships, that is subsumption, identity, paraphrase, elaboration/refinement, etc. For each sentence in the cluster of documents computes: the *centroid score* (a measure of the centrality of a sentence to the overall topic of a cluster), the *position score* (decreases linearly as the sentence gets farther from the beginning of a document), and *overlap-with-first score* (the inner product of the weighted vector representation of the sentence and the first sentence - or title, if there is one). Based on these, produces a *cluster centroid*, consisting of words which are central to all of the documents in the cluster and *ranks sentences* on their distance to the centroid.

Evaluation techniques are very important in text summarization, since it is no obvious how to asses a good summary. One of the most used system used to compare the systems generated summaries to model summaries created manually by professionals, is ROUGE ([4, 5]).
ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) metrics determine n-gram overlaps between system generated summaries and model summaries (human made). A high level of overlap should indicate a high level of shared concepts between the two summaries. Obviously, such metrics are unable to provide any feedback on a summary's coherence.

The *recall, precision and F-measure* used here are defined as usually:

	Relevant	Non-relevant
System relevant	A	B
System non-relevant	C	D

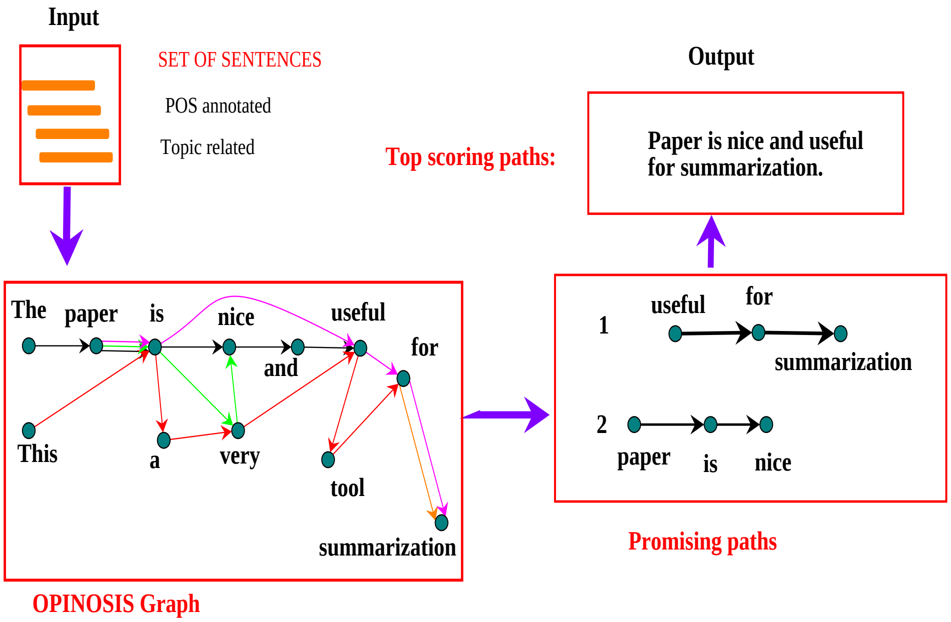
Precision: $P = \frac{A}{A+B}$ **Recall:** $R = \frac{A}{A+C}$ **F-score:** $F = \frac{2PR}{P+R}$

3 A (long) Abstractive Summary of Opinosis

After presenting an example to highlight the approach used in this framework, we give the details of the graph-based approach and, finally, the evaluation machinery managed by the authors is analized.

3.1 The Opinosis Approach Schema

In the figure below, the conceptual schema of Opinosis construction of the summary is described using a simple example.



The system input is a set of POS annotated sentences expressing opinions on the same subject. The **Opinosis Graph** construction module incrementally constructs a graph (starting with thw empty graph) adding the information provided by each sentence. In the (simplified) example above, the first sentence (black) is "The paper is nice and useful" which gives rise to a path which nodes are the words of this sentence, and directed (black) edges expressing the lexical links in the sentence (successive words). The second sentence (green) is a short one "paper is very nice", add to the current graph the new node *very*, and the corresponding (green) edges. Note that the edge *(paper, is)* has been used twice until now, in the graph. The following sentence (magenta) is "paper is useful for summarization" adds three new nodes *useful, for, summarization* and their corresponding lexical links. Finally, the last sentence (red) is "paper is a very useful tool for summarization"; the new nodes *a*, *tool* and (green) edges are added. Clearly, the created edges have different frequencies, and if some shortcuts are permitted, we could also increase them. For example, the edge *(is, nice)* is used only once by the black path, but it could be considered appearing also in the green path if the (unimportant, from the opinion expressed) jumping of the word *very* is permitted. If the jumping of at most two consecutive nodes is permitted, the edge *(is, nice)* could be considered used by all the sentences.

These informations are computed in the **Promising paths** detection module, which constructs a list of most frequently used *sequences of nodes*.

The most promising paths are "composed" to obtain the desired output in the **Top scoring paths** construction and collapsing module.

3.2 The Opinosis-graph and path processing

Graph based representation of texts is frequently used in NLP. The graph constructed in this paper is very simple and it aggregates the individual syntactic links in the whole set of input sentences.

3.2.1 Opinosis-graph construction

Each sentence is tokenized giving rise to a sequence of *word units*, that is a pair $(word, POS\ annotation)$. Words units become *nodes* of OPINOSIS-Graph. The directed edges of the graph are simply pairs (v, w), where v and w are successive word units in the same sentence. To each node v a *Positional Reference Information* list PRI_v is associated, each element of this list beeing a pair (SID_v, PID_v), indicating the index SID_v of the sentence containing v and the position PID_v, of v in this sentence.

The Opinosis-graph $G = (V, E)$ construction is simple: for each (tokenized) input sentence, and for each its word units v, if v is not already a node in the graph, it is added together with its PRI_v initialized on the current pair (index of the current sentence and the position of v in it); if v is already present in th graph, the current pair is added to the list PRI_v; also (excepting the first word unit in each sentence) the lexical link (w, v) is added to the edges of the graph, where w denotes the node preceding v in the current sentence.

The above Opinosis-graph construction intends to

- *capture redundancy*,
- *detect new lexical links*, with the help of "gapped subsequences", and
- *create collapsible structures*, with the help of "hub nodes".

These three goals are realized in the paper by what we call "path processing" and it is described in what follows.

3.2.2 Path processing

A *valid path* in the Opinosis-graph is any sequence of distinct nodes starting from a *valid start node* (vsn), ending in a *valid end node* (ven), satisfying a *gap sequence condition* and a set of *well formed POS constraints*, where:

- a valid start node is any node v s.t. $Average(PID_v) \leq \sigma_{vsn}$, that is, a node which appears frequently in the beginning of the input sentences; σ_{vsn} is a system threshold.

- a valid end node is a punctuation (*period*, *comma*), or any coordinating conjunction (*but*, *yet*).

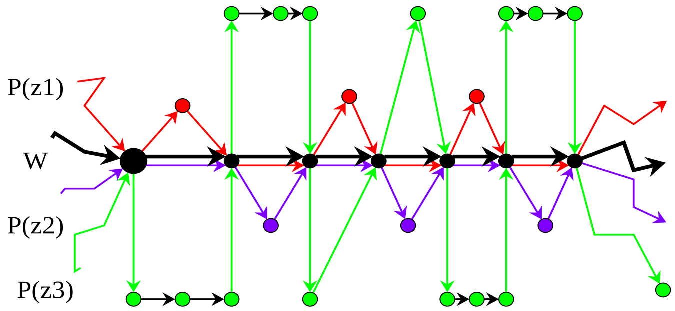
- a well formed POS constraint is specified by a regular-expression constraints in order to satisfy some grammatical rules (for example: $.\star(/\text{nn}) + .\star(/\text{vb}) + .\star(/\text{jj}) + .\star$ or $.\star(/\text{jj}) + .\star(/\text{to}) + .\star(/\text{vb}).\star)$

- the gap sequence condition means that for each two consecutive nodes v and w in the sequence, there is an input sentence in which v and w occurs at a distance not greater than the σ_{gap} threshold of the system.

Let $W = \{v_1, v_2 \ldots, v_s\}$ a sequence of nodes in the Opinosis-Graph. We say that sentence z_i *covers* W if $\forall j \in \{1, \ldots, s-1\}$ $\exists(i,p) \in PRI_{v_j}$ and $\exists(i,p') \in PRI_{v_{j+1}}$ such that $p' - p \leq \sigma_{gap}$. The *path redundancy* of W is :

$$r(W) = |\{z_i | z_i \text{ cover } W\}|.$$

Example. Suppose that the Opinosis-graph was constructed using only three sentences z_1, z_2, z_3 as in the figure bellow.



For $\sigma_{gap} = 1$, the (black) sequence W is not covered by the (red) sentence z_1, since the first pair of nodes of W are not consecutive on z_1. Similarly, the (magenta) sentence z_2 does not cover W because of the second pair of nodes. Also, the (green) sentence z_3 does not cover W since no pair of consecutive nodes is used in w_3. It follows that $r(W) = 0$. However, if we have $\sigma_{gap} = 2$, it is easy to see that both z_1 and z_2 cover W, but z_3 does not cover it. It follows that in this case we have $r(W) = 2$. Finally, taking $\sigma_{gap} = 3$, all three sentences cover W, therefore $r(W) = 3$.

In order to favor a valid path with a high redundancy score, to represent well most of the redundant opinions, the following *path scores* are introduced in Opinosis framework.

If $W = \{v_1, v_2 \ldots, v_s\}$ is a path", then $|W| =$ denotes length of W and $W_{i,j} =$ the subpath $\{v_i, \ldots, v_j\}$ $(1 \leq i < j \leq s)$.

- $S_{basic}(W) = \frac{1}{|W|} \sum_{k=1,s} r(W_{1,k})$
- $S_{wt\_len}(W) = \frac{1}{|W|} \sum_{k=1,s} r(W_{1,k}) \cdot |W_{1,k}|$
- $S_{wt\_loglen}(W) = \frac{1}{|W|} \left[r(W_{1,2}) + \sum_{k=2,s} r(W_{1,k}) \cdot \log |W_{1,k}| \right]$

Clearly, the first score, S_{basic}, favors high redundancy paths; the second, S_{wt_len}, favors moderately redundancy but lengthy paths; the third, S_{wt_loglen} is a trade-off between the first two.

Another interesting path processing introduced in the Opinosis framework is path composition in order to create *stiched sentences*.
A node v_c is called a *collapsible node* if its POS part is *vb*. Any path passing through a collapsible node v_c can be expressed as

$$\overbrace{v_0, \ldots, v_c}^{\text{anchor}},\ \underbrace{v_{first},\ \ \ldots\ ,\ \ v_{last}}_{\text{collapsed candidate}}\ .$$

Then, the set of all collapsed candidates for the node v_c is represented as: $CC(v_c) = \cup_{P\ \text{anchor}} \{P' | P'$ collapsed candidate for $P\}$.

For a fixed anchor P of v_c, let $\{P_1, \ldots, P_{k-1}, P_k\}$ $(k \geq 2)$ be the set of all collapsed candidates for P. The associated *stitched sentence* is:

$$P P_1 \text{comma} P_2 \text{comma} \ldots \text{comma} P_{k-1} \text{cc} P_k.$$

Examples. The sentence "The paper is nice, deep and useful for summarization" could be obtained from the anchor "The paper is" and collapsible candidates "nice", "deep", "useful for summarization". The sentence "The paper is nice, interesting but not useful" could be obtained from the anchor "The paper is" and collapsible candidates "nice", "interesting ", "not useful".

The algorithm suggested in the paper for choosing the coordinating conjunction before the last collapsed candidate is: *from all predecessors u in G of the first node v of P_k, having POS=cc, select* $\text{argmax}_{u:POS(u)=cc} r(\{u, v\})$.

3.3 The Opinosis overall algorithm

The *system's parameters* used in the algorithm (their values were set experimentally) are:

- σ_{gap} - controls the maximum allowed gaps in discovering redundancies.
- σ_{vsn} - qualify nodes that tend to occur early on in a sentence.
- σ_{ss} - controls the maximum number of winning paths (summary size).
- σ_r - a redundancy score threshold, to prune non-promising paths.

The algorithm main phases and their description: *Starting with a set* $Z = \{z_i\}_{i=1}^{n}$ *of topic related sentences to be summarized, the algorithm outputs* $\mathcal{O} = \{Opinosis\ Summary\}$ as follows

1. Construct the Opinosis-graph G from Z.
2. For each node v of G, qualified as a valid start node, execute a recursive Depth First Search to find *valid paths*.
   (a) The PRI overlap information, path length and score are maintained during the search. σ_r is used to avoid paths over generation.
   (b) A list $\mathcal{C}$ of promising sentences is maintained.
   (c) When a collapsible node v_c is reached, the corresponding collapsed candidates (obtained when returning in v_c) are composed with the current path to v_c and the stitched sentence is added to $\mathcal{C}$.
3. Sort $\mathcal{C}$ non increasing by *path scores* and return $\mathcal{O}$, containing the first σ_{ss} sentences of $\mathcal{C}$.

3.4 System evaluation

For the experimental setup, real data have been considered, namely reviews from specialized sites as `Tripadvisor.com, Amazon.com`, and `Edmunds.com`. It were selected 51 review documents each about an entity E and a topic X. Their size was about 100 sentencens per review document. For each review document the best 4 human reference (realized by 5 professional reviewers) summaries were considerated.

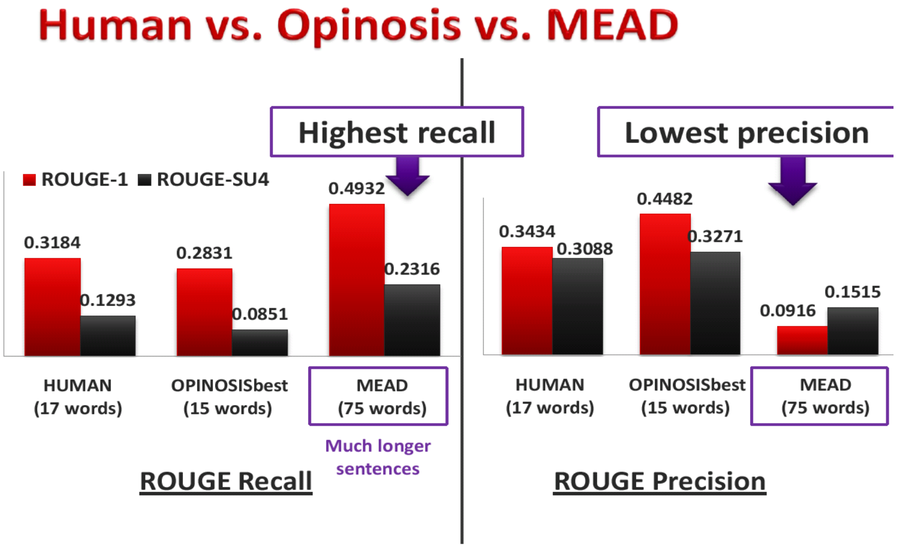
Performance comparison has been made between *humans*, *Opinosis* and the *baseline method* **MEAD**. For each reference summary it was computed ROUGE scores over the remaining 4-1=3 reference summaries. Method

MEAD selects 2 most representative sentences as summaries. Corresponding, the Opinosis parameter σ_{ss} were set to 2. Also, in comparison Opinosis used the best setting of the other parameters ($\sigma_{vsn} = 15$, $\sigma_{gap} = 2$, $\sigma_r = 4$) and the S_{wt_loglen} score. **ROUGE scores** were reported with the use of stemming and stopword removal.

An interesting **readability test** is proposed: *mix N sentences from the system summary $\mathcal{O}$ and M sentences from human summary, and ask a human assessor to pick at most N sentences that are least readable. Then,*

$$readability(\mathcal{O}) = 1 - \frac{\#CorrectPick}{N}.$$

The following figure summarizes the results of the performance comparison:

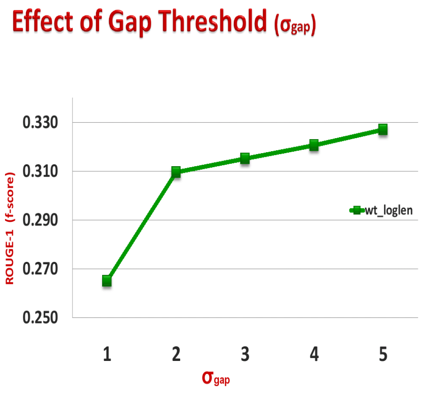


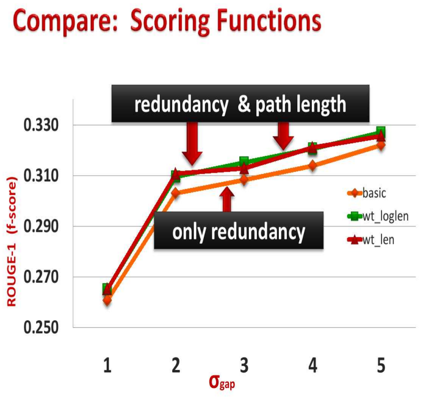
The very high recall scores and extremely low precision scores obtained by MEAD are explained by the extractive method used in this system.

It could observe reasonable agreement amongst humans, their results being better than Opinosis but comparable to MEAD.

Opinosis is closer in performance to humans than to Mead. The recall scores of Opinosis summaries are slightly lower than that achieved by humans, while the precision scores are higher. The improvement of precision by Opinosis over that of humans is more significant than the decrease of recall (Wilcoxon test).

Also, the experiments revealed the best σ_{gap} parameter and the best score, as depicted in the following figure:





Readability test. Human assessor picked the least 102 readable from 565 sentences (102 were Opinosis generated). Out of these 102 , the human assessor picked only 34, resulting in an *average readability score of 0.67.* From the 34 sentences with problems generated by Opinosis: 11 contained no information, incomprehensible; 12 were incomplete (false positives of validity check); 8 had conflicting information (e.g. ”the hotel room is clean and dirty”); 3 were considered having ”poor grammar”.

4 Summary of Our Opinions on Opinosis

Positive Opinions:

- The paper introduces a nice, interesting and simple approach to abstractive summarization. It is based only by frequencies of lexical links (order of words in the input text). Discovering that for highly redundant corpora this is a good solution is important and it has not been used in the NLP realm.

- The new approach, carrying ideas from sequential data mining, could be a very useful instrument in business decision taking based on the (quantitative) results of consumer surveys. In fact, the last author of this paper is one of prominent scientist in sequential data mining.

- The evaluation setup of the system is impressive; in fact, for the summarization field, evaluation methods and strategies are very important, due to the strong subjective quality of the human reference.

- Finally, two interesting research ideas suggested by the analysis of this paper:

  - Path-aggregation building of a (social) network. Usually, social networks are constructed by adding a new node to an existing network. We have here a real example in which the network is incrementally constructed by paths! This could be interesting, for example, in building cooperative maps of mobile agents.
  - Adding OPINOSIS to *conference management software systems* (e.g. EasyChair)? Of course, a summary of the reviewers for the accepted papers (or for the rejected ones) at a big conference could be of real interest for the respective field of research.

Not quite positive Opinions:

- Gapping could change an opinion. Jumping a negation in a sentence is quite dangerous. For example, the above title could be considered as ”positive Opinions”, which is not true. It follows that a mechanism to control the word units jumped must be added. The PRI lists management must be changed correspondingly.

- Detection of synonyms, an usual shallow NLP tool, is not considered, despite of their obvious influence in the redundancy scores. This could be done by extending the existence test of a node in the graph (a node already exists, if one of its synonyms with the same POS is present).

- Emphasizes too much on the surface order of words (the stitched sentences are not quite abstractive sentences; they are not related to the most important words in the text).

- In the algorithm for constructing the stitched sentence, considering only the last determined collapsible candidate responsible for the coordinating conjunction added, is not justified. Any collapsible candidate can be the last one and therefore the algorithm for determining the coordinating conjunction must consider all the first unit words of all candidates. In fact, a high importance on generating correct sentences would be a (weak) partition of the collapsible candidates in two classes expressing positive opinions, respectively negative opinions. Taking in account the corresponding two types of attributes seems necessary. This is, clearly, a weak point of the paper.

- Shallow NLP must be reward by learning. If no other NLP tools are considered, then must be tested some other more sofisticated inference methods than the straightforward sorting of promising paths on the scores (see, for example, [2]).

- We believe that in the construction of the Opinosis graph it is necessary to take into consideration (as in evaluation experiments) the removal of stop words, which would lead to more accurate results.

- There are some mathematical inaccuracies in the text (for example, in the definition of the redundancy which is a number but in the text is defined as a set).

- Algorithms are not optimized for huge corpus. For example, the use of PRI lists and of the gap threshold could be improved if an augmented graph (considering the gap induced edges) is firstly created.

- Last, but not least, the authors do not mention in the paper the cases when their rules could generate inaccuracies. They just take some examples of well formed sentences which are perfectly for their cases, but in real applications, when we have a lots of different opinions, the results could be very unpleasant. The tuning of the various parameters of the system does not help.

References

[1] DiMarco Chrysanne, Hirst Graeme, Hovy Eduard, *Generation by selection and repair as a method for adapting text for the individual reader*, Proceedings, Workshop on Flexible Hypertext, Eighth ACM International Hypertext Conference, Southampton, U.K., April 1997.

[2] Filippova Katja, *Multi-sentence compression: Finding shortest paths in word graphs*, Proceedings of the 23rd International Conference on Computaional Linguistics (COLING 10), Beijing, China, pp. 322-330, 2010

[3] Ganesan Kavita, Zhai ChengXiang and Han Jiawei, *"Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions"*, Proceedings of the 23rd International Conference on Computational Linguistics (COLING 10), Beijing, China, pp. 340-348, 2010

[4] Lin Chin-Yew, *Looking for a few good metrics : Rouge and its evaluation*, Proceedings of the 4th NTCIR Workshops, 2004

[5] Lin Chin-Yew, *Rouge: a package for automatic evaluation of summaries*, Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 2004

[6] McKeown Kathleen, Klavans Judith, Hatzivassiloglou Vasileios, Barzilay Regina, Eskin Eleazar, *Towards Multidocument Summarization by Reformulation: Progress and Prospects*, AAAI/IAAI 453-460, 1999

[7] Open NLP, `http://opennlp.sourceforge.net./`

[8] Radev Dragomir, Hongyan Jing, and Malgorzata Budzikowska, *Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies* , ANLP/NAACL Workshop on Summarization, pages 21-29, 2000.

[9] Radev Dragomir, Hovy H. Eduard, McKeown Kathleen, *Introduction to the special issue on summarization*, Computational Linguistics 28(4): 399-408 (2002)