Multi-document Summarization Using Bipartite Graphs

Daraksha Parveen and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
(daraksha.parveen|michael.strube)@h-its.org

Abstract

In this paper, we introduce a novel graph based technique for topic based multi-document summarization. We transform documents into a bipartite graph where one set of nodes represents entities and the other set of nodes represents sentences. To obtain the summary we apply a ranking technique to the bipartite graph which is followed by an optimization step. We test the performance of our method on several DUC datasets and compare it to the state-of-the-art.

1 Introduction

Topic-based multi-document summarization aims to create a single summary from a set of given documents while considering the topic of interest. The input documents can be created by querying an information retrieval or search engine for a particular topic and retaining highly ranked documents, or by clustering documents of a large collection and then using each cluster as a set of input documents (Galanis et al., 2012). Here, each cluster of the set of documents contains a representative topic.

A summary extracted from a set of input documents must be related to the topic of that set. If textual units (or sentences) extracted from different documents convey the same information, then those units are called redundant. Ideally, the multi-document summary should be non-redundant. Hence each textual unit in a summary should convey unique information. Still, all extracted textual units should be related to the topic. They should also make up a coherent summary.

When building summaries from multiple documents belonging to different sets, a system should attempt to optimize these three basic properties:

1. **Relevance**: A summary should contain only those textual units which are relevant to the topic and provide useful information.
2. **Non-redundancy**: A summary should not contain the same information twice.
3. **Readability**: A summary should have good readability (syntactically well formed, no dangling pronouns, coherent, ...).

Generally, multi-document summarization systems differ from each other on the basis of document representation, sentence selection method or on the requirements for the output summary. Popular methods for document representation include graph-based representations (e.g. *LexRank* (Erkan and Radev, 2004) and *TextRank* (Mihalcea and Tarau, 2004)) and tf-idf vector-based representations (Luhn, 1958; Nenkova and Vanderwende, 2005; Goldstein et al., 2000). These document representations act as input for the next phase and provide information about the importance of individual sentences. Sentence selection is the crucial phase of the summarizer where sentence redundancy must be handled in an efficient way. A widely used technique is the greedy approach introduced by Carbonell and Goldstein (1998) and Goldstein et al. (2000). They compute a relevance score for all sentences with regard to the topic, start by extracting the most relevant sentence, and then iteratively extract further sentences which are relevant to the topic and at the same time most dissimilar to already extracted sentences. Later more fundamental optimization methods have been widely used in multi-document summarization, e.g. Integer Linear Programming (ILP) (McDonald, 2007; Gillick et al., 2009; Nishikawa et al., 2010; Galanis et al., 2012). Unlike most other approaches (Galanis et al., 2012) has also taken into account the readability of the final summary.

In this work, we introduce an extractive topic based multi-document summarization system which represents documents graphically and

Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing, pages 15–24,
October 29, 2014, Doha, Qatar.

optimizes the importance of sentences and non-redundancy. The importance of sentences is obtained by means of applying the Hubs and Authorities ranking algorithm (Kleinberg, 1999) on the unweighted bipartite graph whereas redundancy in the final summary is dealt with entities in a graph.

In Section 2 we introduce the state-of-the-art in topic based multi-document summarizaton. Section 3 provides a detailed description of our approach. Experiments are described in Section 4 where we also briefly describe the datasets used and the results. Section 5 discusses the results of our approach, and in Section 6 we finally give conclusions.

2 Related work

A graph-based representation of documents for summarization is adopted by various approaches. For instance, *TextRank* by Mihalcea and Tarau (2004) applies the *PageRank* algorithm (Brin and Page, 1998) to extract important sentences for single document summarization. This ranking algorithm proclaims the importance of a sentence by considering the global information which is computed recursively from the entire graph. Later, the graph is converted into a weighted graph in which the weights are calculated by measuring the similarity of sentences (Mihalcea, 2004). Similarly, in the *LexRank* approach (Erkan and Radev, 2004), documents are represented as a similarity graph in which the sentences are nodes and these sentences are then ranked according to centrality measures. The three centrality measures used are degree, *LexRank* with threshold and continuous *LexRank*. *LexRank* is a measure to calculate ranks using the similarity graph of sentences. It is also known as lexical *PageRank*. The summarization approach developed by Gong and Liu (2001) is also based on ranking sentences where important sentences are selected using a relevance measure and latent semantic analysis.

Later, for better performance, sentences are classified according to their existence in their final summary in binary format i.e. 1 (belongs to summary) and 0 (doesn't belong to summary) (Shen et al., 2007; Gong and Liu, 2001). Here, the sentences are projected as feature vectors and conditional random fields are used to classify them. During document processing, most informative sentences are selected by the summarizer (Shen et al., 2007). Fattah and Ren (2009) also considers summarization as two class classification problem. They use a genetic algorithm and mathematical regression to select appropriate weights for the features and used different classification technique for e.g. feed forward neural network, probablistic neural network and Gaussian mixture models.

In the summarization task, optimization of the three properties discussed in Section 1, relevance, non-redundancy and readability, is required. This is a global inference problem, which can be solved by two approaches. Firstly, relevance and redundancy can be optimized simultaneously. For instance, Goldstein et al. (2000) developed a metric named MMR-MD (influenced by the Maximum Marginal Relevance (MMR) approach of Carbonell and Goldstein (1998)) and applied it to clusters of passages. Similarly, influenced by the SumBasic system (Nenkova and Vanderwende, 2005), Yih et al. (2007) developed a system which assigns a score to each term on the basis of position and frequency information and selects the sentence having highest score. Other approaches are based on an estimate of word importance (e.g. Lin and Hovy (2000)) or the log likelihood ratio test which identifies the importance of words using a supervised model that considers a rich set of features (Hong and Nenkova, 2014). Finally, Barzilay and Elhadad (1999) extract sentences which are strongly connected by lexical chains for summarization. The second approach deals with relevance and redundancy seperately. For instance, McKeown et al. (1999) create clusters of similar sentences and pick the representative one from every cluster. The representative sentence of a cluster of sentences takes care of the requirement to extract relevant information whereas clustering reduces the redundancy.

McDonald (2007) proposes a new ILP optimization method for extractive summarization. He introduces an objective function which maximizes the importance of sentences and minimizes the similarity of sentences. ILP methods for optimization have also been adopted by Berg-Kirkpatrick et al. (2011),Woodsend and Lapata (2012) and Galanis et al. (2012). Until now, Galanis et al. (2012) have reported the highest scores for multi-document summarization on DUC2005 and DUC2007. However, their approach is not completely unsupervised.

3 Our method

This section describes the technique, which we adopted for summarization. We start by discussing the graphical representation of the text followed by a description how to quantify the importance of sentences in the input texts. We then discuss the ILP technique which optimizes the importance of sentences and redundancy.

3.1 Graphical representation of text

The graphical representation of a text makes it more expressive than a traditional *tf-idf* depiction for summarization. A graph can easily capture the essence of the whole text without leading to high computational complexity. Guinaudeau and Strube (2013) introduced a bipartite graph representation of text based on the entity grid (Barzilay and Lapata, 2008) representation of text. The projection of this bipartite graph representation has been used for calculating the local coherence of a text (Guinaudeau and Strube, 2013). The basic intuition to use a bipartite graph for summarization is that it contains entity transitions similar to lexical chains (Barzilay and Elhadad, 1999). An appropriate measure to determine the importance of sentences by considering strong entity transitions indicates the information central to a text better than simply giving scores on the basis of most frequent words. The unweighted bipartite graph $G = (V_s, V_e, L)$ contains two sets of nodes, V_s corresponding to the sentences from the input text and V_e corresonding to the entities, and a set of edges represented by L. Figure 1 shows a model summary from the DUC 2006 data, which is transformed into an entity grid in Figure 2 (Barzilay and Lapata, 2008; Elsner and Charniak, 2011). Here, cells are filled with the syntactic role a mention of an entity occupies in a sentence. Subjects are denoted by *S*, objects by *O* and all other roles by *X*. If an entity is not mentioned in a sentence then the corresponding cell contains "-". In the corresponding bipartite graph (Figure 3), edges are created between a sentence and an entity only if the entity is mentioned in a sentence (the cell in entity grid is not "-"). Since this is a dyadic graph, there are no edges between nodes of the same set.

3.2 Ranking the importance of sentences

A graph based ranking algorithm is used to calculate the importance of a sentence represented as a node in the graph discussed above. In contrast to the local information specific to a vertex, graphical ranking algorithms take (graph-) global information to calculate the rank of a node. The *Hyperlink-Induced Topic Search* algorithm (*HITS*, also known as *Hubs and Authorities*) by Kleinberg (1999) is used to rank sentences in our method. This algorithm considers two types of nodes, hence it is well suited to rank sentences in our bipartite graph. Entities are considered as hub nodes, and sentences are considered as authority nodes. The importance of a sentence is calculated in two steps:

- Hub update rule: Update each node's hub score to be equal to the sum of the authority scores of each node that it points to. It can be written as:

$$HubScore = A \cdot AuthorityScore \quad (1)$$

Here, A is an adjacency matrix which represents the connection between the nodes in a graph.

- Authority update rule: In this step, each authority node is updated by equating them to the sum of the hub scores of each node, which is pointing to that authority node. It can be written as:

$$AuthorityScore = A^T \cdot HubScore \quad (2)$$

Hence, the authority weight is high if it is pointed at by a hub having high weights.

Given some intial ranks to all nodes in a graph, the hub and authority update rules are applied until convergence. After applying this algorithm, the rank of every node is obtained. The rank is considered as importance of the node within the graph. We normalize the ranks of sentences according to sentence length to avoid assigning high ranks to long sentences.

To incorporate important information from documents, ranks of entities are incremented by $Rank + tf_{doc} \cdot idf_{doc}$ in every iteration, where tf_{doc} shows the importance of an entity in a document by calculating the frequency whereas idf_{doc} is an inverse document frequency from the current cluster. $Rank + tf_{doc} \cdot idf_{doc}$ is used in calculating the AuthorityScore. Initially, the $Rank$ can be any numerical value but after every iteration of the HITS algorithm it will be updated accordingly.

S_1 The treatment of osteoarthritis includes a number of non-steroidal anti-inflammatory drugs such as aspirin, acetaminophen, and ibuprofen.

S_2 These drugs, however, cause liver damage and gastrointestinal bleeding and contribute to thousands of hospitalizations and deaths per year.

S_3 New cox-2 inhibitor drugs are proven as effective against pain, with fewer gastrointestinal side effects.

S_4 The two together appeared to reduce knee pain after 8 weeks.

Figure 1: Model summary from DUC 2006

	TREATMENT (e1)	OSTEOARTHRITIS (e2)	NUMBER (e3)	DRUGS (e4)	ASPIRIN (e5)	ACETAMINOPHEN(e6)	IBUPROFEN (e7)	DAMAGE (e8)	BLEEDING (e9)	THOUSANDS (e10)	DEATHS (e11)	YEAR (e12)	PAIN (e13)	EFFECTS (e14)	TWO (e15)	WEEKS (e16)
S_1	S	X	O	X	X	X	X	-	-	-	-	-	-	-	-	-
S_2	-	-	-	S	-	-	-	O	O	X	X	X	-	-	-	-
S_3	-	-	-	S	-	-	-	-	-	-	-	-	X	X	-	-
S_4	-	-	-	-	-	-	-	-	-	-	-	-	O	-	S	X

Figure 2: Entity grid of the model summary from Figure 1

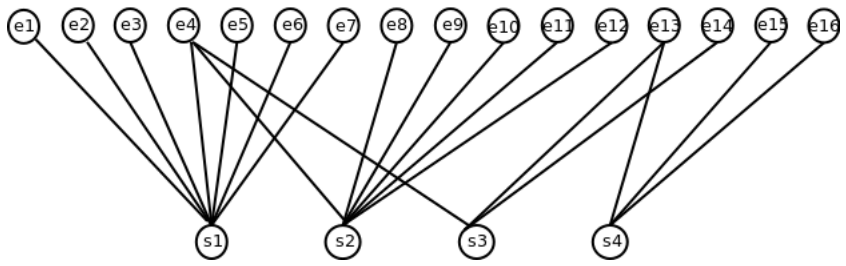

Figure 3: Bipartite graph derived from the entity grid from Figure 2

3.3 Optimization algorithm

In topic-based multi-document summarization, the final summary should be non-redundant. At the same time it should contain the important information from the documents. To achieve these two conditions, we employ integer linear programming (ILP) to obtain an optimal solution. In ILP we maximize an objective function. Our objective function, given in Equation 3, has two parts: the importance of a summary and the non-redundancy of a summary. The values obtained after ranking by the HITS algorithm are used as the importance of sentences for ILP. Non-redundancy can not be calculated for a single sentence. Instead, it has to be evaluated with respect to other sentences. We calculate non-redundancy by the number of unshared entities, i.e. entities which are not shared by other sentences, after appending a sentence to a summary. The least redundant sentence will increase the number of entities in the final summary.

$$max(\lambda_1 \sum_{i=1}^{n} (Rank(s_i) + topicsim(s_i)) \cdot x_i + \lambda_2 \sum_{j=1}^{m} y_j) \quad (3)$$

Equation 3 is the objective function where m is

	Topic	Documents per topic	Human Summaries	Word limit in final summary
DUC 2005	50	25-50	4-9	250
DUC 2006	50	25	4	250
DUC 2007	45	25	4	250

Table 1: Document Statistics

the number of entities in a document and n is the number of sentences in a document. x_i and y_j are binary variables for sentences and entities respectively. λ_1 and λ_2 are tuning parameters. $Rank(s_i)$ is a rank of a sentence s_i obtained by applying the HITS algorithm. Since, we work on topic-based multi-document summarization, we include topic information by calculating $topicsim(s_i)$, which captures the cosine similarity of a sentence s_i with the corresponding topic. If the topic contains more than one sentence then we take an average of cosine similarity with a sentence s_i. The constraints on the variables are shown in Equations 4-6:

$$\sum_{i=1}^{n} Len(s_i) \cdot x_i \leq Len(summary) \quad (4)$$

Here, $Len(s_i)$ and $Len(summary)$ are the number of words in a sentence s_i and in the final summary, respectively. This constraint does not allow the length of final summary to exceed its maximum length. The maximum length varies depending on the datasets discussed in Section 4.1.

$$\sum_{j \epsilon E_i} y_j \geq Entities(s_i), \text{for } i = 1, \ldots, n \quad (5)$$

In constraint 5, E_i is a set of entities present in a sentence s_i. The number of entities present in a sentence is represented as $Entities(s_i)$. If a sentence s_i is selected then the entities present in a sentence are also selected($\sum y_j = Entities(s_i)$). Whereas, if a sentence s_i is not selected then some of its entities can also be selected because they may appear in already selected sentences ($Entities(s_i) = 0, \therefore \sum y_j \geq 0$). In both the cases, constraint 5 is not violated.

$$\sum_{i \epsilon S_j} x_i \geq y_j, for j = 1, \ldots, m \quad (6)$$

In constraint 6, S_j is a set of sentences containing entity y_j. This constraint shows that, if an entity y_j is selected then at least one sentence is selected which contains it ($y_j = 1, \therefore \sum x_i \geq 1$). If an entity y_j is not selected, then it is possible that none of the sentences which contain it may not be selected ($y_j = 0, \therefore \sum x_i = 0$). Also, constraint 4 holds in either of the cases.

4 Experiments

We perform experiments on various DUC datasets to compare the results with state-of-the-art systems.

4.1 Datasets

Datasets used for our experiments are DUC2005 (Dang, 2005), DUC2006 (Dang, 2006) and DUC2007[1] . Each dataset contains group of related documents. Each group of documents contains one related topic or a query consisting of a few sentences. In DUC, the final summary should respond to the corresponding topic. Also, the summary cannot exceed the maximum allowed length. For instance, in DUC2005, 250 words are allowed in the final summary. Every document cluster has corresponding human summaries for evaluating system summaries on the basis of ROUGE scores (Lin, 2004). The sources of DUC datasets are Los Angeles Times, Financial Times of London, Associated Press, New York Times and Xinhua news agency. We employ ROUGE SU4 and ROUGE 2 as evaluation metrics. ROUGE returns recall, precision and F-score of a system, but usually only recall is used in for evaluating automatic summarization systems, because the final summary does not contain many words. Hence, if the recall is high then the summarization system is working well. Document statistics is provided in Table 1.

4.2 Experimental setup

We use raw documents from the various DUC datasets as input for our system. We remove non-alphabetical characters from the documents. Then we obtain a clean sentence split by means of the Stanford parser (Klein and Manning, 2003) so that the sentences are compatible with the next steps.

[1] http://www-nlpir.nist.gov/projects/duc/index.html

	ROUGE-2	ROUGE-SU4
$\lambda_1 = 0.5$ & $\lambda_2 = 0.5$	0.07950	0.14060
$\lambda_1 = 0.6$ & $\lambda_2 = 0.4$	0.07956	0.14071
$\lambda_1 = 0.7$ & $\lambda_2 = 0.3$	0.07975	0.14105
$\lambda_1 = 0.8$ & $\lambda_2 = 0.2$	0.07976	0.14106
$\lambda_1 = 0.9$ & $\lambda_2 = 0.1$	0.07985	0.14107

Table 2: Results on different λ's on DUC 2005

Dataset	ROUGE-2	ROUGE-SU4
DUC 2005 (32)	0.07975 (1)	0.14105 (1)
DUC 2006 (35)	0.08969 (3)	0.15070 (2)
DUC 2007 (32)	0.10928 (6)	0.16735 (5)

Table 3: System performance (and rank) on the DUC 2005, 2006 and 2007 (main) data. The number in parenthesis after the DUC year indicates the number of competing systems.

We use the Brown coherence toolkit (Elsner and Charniak, 2011) to convert the documents into the entity grid representation from which the bipartite graph is constructed (Guinaudeau and Strube, 2013). Entities in the graph correspond to head nouns of noun phrase mentioned in the sentences. The ranking algorithm from Section 3.2 is applied to this graph and returns the importance score of a sentence as required by the objective function given in Equation 3. Next optimization using ILP is performed as described in Section 3.3. We use GUROBI Optimizer[2]for performing ILP. ILP returns a binary value, i.e., if a sentence should be included in the summary it returns 1, if not it returns 0. We set $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$ for all datasets. We did not choose the optimal values, but rather opted for ones which favor importance over non-redundancy. We did not observe significant differences between different λ values as long as $\lambda_1 > \lambda_2$ (see Table 2). The sentences in the output summary are ordered according to their ranks. If the output summary contains pronouns, we perform pronoun resolution in the source documents using the coreference resolution system by Martschat (2013). If pronoun and antecedent occur in the same sentence, we leave the pronoun. If the antecedent occurs in an earlier sentence, we replace the pronoun in the summary by the first element of the coreference chain the pronoun belongs to. Except for setting λ_1 and λ_2 on DUC 2005, our approach is unsupervised, as there is no traning data required. The recall (ROUGE) scores on different datasets are shown in Table 3.

Table 3 shows that our system would have performed very well in the DUC 2005 and DUC 2006 competitions with ranks in the top 3 and well in the DUC 2007 competition. Since the competitions date a while back, we compare in addition to the current state-of-art in multi-document summarization. To our knowledge Galanis et al. (2012) report the best results on DUC 2005 data. While their ROUGE-2 score is slightly better than ours, we outperform them in terms of ROUGE-SU4 (0.14105 vs. 0.13640), where, to our knowledge, our results are the highest reported so far. However, their results on DUC 2007 (ROUGE-2 0.12517 and ROUGE-SU4 0.17603) are still quite a bit better than our results. On the DUC 2006 data we outperform the HIERSUM system by Haghighi and Vanderwende (2009) on ROUGE-2 (0.08969 vs. 0.086) as well as on ROUGE-SU4 (0.15070 vs. 0.143). On the DUC 2007 data, our results are worse than theirs on ROUGE-2 (0.10928 vs. 0.118) and on par on ROUGE-SU4 (0.16735 vs. 0.167). The system which won the DUC 2007 task, PYTHY by Toutanova et al. (2007), performs similar to HIERSUM and hence slightly better than our system on these data. The recent work by Suzuki and Fukumoto (2014) evaluates also on DUC 2007 but reports only ROUGE-1 scores. We obtain a ROUGE-1 score of 0.448 on DUC 2007 which is better than Suzuki and Fukumoto (2014) (0.438) as well as PYTHY (0.426). The best ROUGE-1 score reported to date has been reported by Celikyilmaz and Hakkani-Tür (2010) with 0.456. The difference between this score and our score of 0.448 is rather small.

5 Discussion

Several approaches have been proposed for topic based multi-document summarization on the DUC datasets we use for our experiments. The best results to date have been obtained by supervised and semi-supervised systems. The results of our system are mostly on par with these systems though our system is unsupervised (as mentioned in Section 4 the values for λ_1 and λ_2 in the objective function (Equation 3) were not tuned for optimal ROUGE scores but rather set for favoring importance over non-redundancy).

We compared our results with various state-of-

[2]Gurobi Optimization, Inc., `http://www.gurobi.com`

S_1 What is being learned from the study of deep water, seabeds, and deep water life?

S_2 What equipment and techniques are used?

S_3 What are plans for future related activity?

Figure 4: Topic containing interrogative words from DUC 2007

S_1 I've started to use irrigation hoses called "leaky pipe".

S_2 Soil's usually best to water the target area a few days before I plan to dig.

S_3 If I don't place element in the root zone , element can't be added later when the plants are growing.

S_4 The new composts were much lighter and more suitable for container plants in garden centres and through these were rapidly introduced to gardeners.

Figure 5: Sentences containing dangling first person pronoun from DUC 2005

the-art systems, and our system is giving competitive results in both ROUGE-2 and ROUGE-SU4 scores. However, the ROUGE-2 score of Galanis et al. (2012) on DUC 2005 is slightly better than our score. This might be because they use bigram information for redundancy reduction. However, they need training data for sentence importance. Hence their system has to be classified as supervised while ours is unsupervised.

We have also calculated the ROUGE-1 score on DUC 2007 and compared it with state-of-the-art approaches. HybHsum (Celikyilmaz and Hakkani-Tür, 2010) has obtained the top ROUGE-1 score on DUC 2007 with 0.456. However, HybHsum is a semi-supervised approach which requires a labeled training data. The difference between our ROUGE-1 score of 0.448 and HybHsum ROUGE-1 score on DUC2007 is not significant (to be fair, achieving significant improvements in ROUGE scores on DUC data is very difficult). In contrast to HybHsum, our approach is unsupervised.

Our method computes importance on the basis of a bipartite graph. We believe that our bipartite graph captures more information than the general graphs used in earlier graph-based approaches to automatic summarization. Entity transition information present in the bipartite graph of a document, helps us in finding the salient sentences. Our approach works well if the graph is not sparse.

We observed a couple of problems in the output of our system which we plan to address in future work. If topics contain interrogative pronouns as shown in Figure 4 the mapping between topic and sentences from the documents does not work well. We need to resolve which entities the interrogative pronouns refer to. Another problem occurs, because the coreference resolution system employed does not resolve first person pronouns. Hence, we end up with summaries containing dangling first person pronouns as shown in Figure 5. However, our system appears to work reasonably well in other cases where the summaries are coherent and readable and also have a high ROUGE score as shown in the summary from DUC 2007 data in Figure 6.

6 Conclusions

In this paper, we have presented an unsupervised graph based approach for topic based multi-document summarization. Our graph based approach provides state-of-the-art results on various datasets taken from DUC competitions. The graph based representation of a document makes computation very efficient and less complex. In future work, we incorporate the syntactic roles of entities, to provide more information in the method.

Acknowledgments

This work has been funded by Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship.

The European Parliament , angered by Turkey 's human rights record , voted Thursday to freeze hundreds of millions of US dollars in aid to Turkey for setting up a customs union with the EU. Since then , the EU has been trying to patch up the relationship , with several leaders of member countries insisting that Turkey 's place is in the union. The special aid is part of the agreement between the European Union EU and Turkey on the establishment of a customs union between the two sides. " The European Union , without renouncing its principles , " will have to decide in December to allow Turkey to become a formal candidate for EU membership. ANKARA , February 27 Xinhua Turkey today welcomed the European Union 's attitude toward its dispute with Greece and urged the EU to release financial assistance immediately despite Greek efforts to block it. After the decision in December to exclude Turkey from the first wave of enlargement talks , Turkey put its relations with the 15 member union on hold. During Solana stay here , Turkish leaders reiterated their position to link the expansion of the NATO with Turkey 's entry into the European Union. The European Union , European Union Ankara wants to join , is pressing Turkey to find a peaceful solution to the war. The statement added that Greece , despite its attempts , was unable to get the support of the other 14 European Union members in getting a statement that would express solidarity with Greece and condemn Turkey. Both the European Union and the United States criticized Turkey for jailing Birdal.

Figure 6: Output summary from DUC 2007

Acknowledgments

This work has been funded by Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship.

References

Regina Barzilay and Michael Elhadad. 1999. Using lexical chains for text summarization. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 111–121. Cambridge, Mass.: MIT Press.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pages 481–490.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, 24–28 August 1998, pages 335–336.

Asli Celikyilmaz and Dilek Hakkani-Tür. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 815–824.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Conference held at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 9–10 October 2005.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Proceedings of the 2006 Document Understanding Conference held at the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* New York, N.Y., 8–9 June 2006.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the ACL 2011 Conference Short Papers,* Portland, Oreg., 19–24 June 2011, pages 125–129.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Mohamed Abdel Fattah and Fuji Ren. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech and Language*, 23(1):126–144.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pages 911–926.

Daniel Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2009. A global optimization framework for meeting summarization. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing,* Taipei, Taiwan, 19–24 June 2009, pages 4769–4772.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the Workshop on Automatic Summarization at ANLP/NAACL 2000,* Seattle, Wash., 30 April 2000, pages 40–48.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* New Orleans, Louis., 9–12 September 2001, pages 19–25.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* Sofia, Bulgaria, 4–9 August 2013, pages 93–103.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 362–370.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Sweden, 26–30 April 2014, pages 712–721.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pages 423–430.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for automatic summarization. In *Proceedings of the 18th International Conference on Computational Linguistics,* Saarbrücken, Germany, 31 July – 4 August 2000, pages 495–501.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out Workshop at ACL '04,* Barcelona, Spain, 25–26 July 2004, pages 74–81.

H.P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165.

Sebastian Martschat. 2013. Multigraph clustering for unsupervised coreference resolution. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics,* Sofia, Bulgaria, 5–7 August 2013, pages 81–88.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the European Conference on Information Retrieval,* Rome, Italy, 2-5 April 2007.

Kathleen R. McKeown, Judith L. Klavans, Vassileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence,* Orlando, Flo., 18–22 July 1999, pages 453–460.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, pages 404–411.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 170–173.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical Report MSR-TR-2005-101, Microsoft Research.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 910–918.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January 2007, pages 2862–2867.

Yoshimi Suzuki and Fumiyo Fukumoto. 2014. Detection of topic and its extrinsic evaluation through multi-document summarization. In *Proceedings of the ACL 2014 Conference Short Papers,* Baltimore, Md., 22–27 June 2014, pages 241–246.

Kristina Toutanova, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamudi, Hisami Suzuki, and Lucy Vanderwende. 2007. The PYTHY summarization system: Microsoft Research at DUC 2007.

In *Proceedings of the 2007 Document Understanding Conference held at the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 26–27 April 2007.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 233–242.

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January 2007, pages 1776–1782.