CTSUM: Extracting More Certain Summaries for News Articles

Xiaojun Wan and Jianmin Zhang*
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistics (Peking University), MOE, China
wanxiaojun@pku.edu.cn, 540051364@qq.com

ABSTRACT
People often read summaries of news articles in order to get reliable information about an event or a topic. However, the information expressed in news articles is not always certain, and some sentences contain uncertain information about the event. Existing summarization systems do not consider whether a sentence in news articles is certain or not. In this paper, we propose a novel system called CTSUM to incorporate the new factor of information certainty into the summarization task. We first analyze the sentences in news articles and automatically predict the certainty levels of sentences by using the support vector regression method with a few useful features. The predicted certainty scores are then incorporated into a summarization system with a graph-based ranking algorithm. Experimental results on a manually labeled dataset verify the effectiveness of the sentence certainty prediction technique, and experimental results on the DUC2007 dataset shows that our new summarization system cannot only produce summaries with better content quality, but also produce summaries with higher certainty.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*

General Terms
Algorithms, Performance, Design, Human Factors.

Keywords
CTSUM; information certainty; multi-document summarization.

1. INTRODUCTION
With the rapid growth of on-line digital content publishing and propagation, there are usually hundreds or thousands of news articles about an event or topic. People often read summaries of news articles in order to obtain useful and certain information about an event or a topic. Existing multi-document summarization systems (e.g. NewsInEssence [28], NewsBlaster [7]) can be used to produce a short summary for a collection of news articles.

*Jianmin Zhang is an undergraduate student in School of Information Science and Technology, Beijing Normal University. The work was done while she was an intern in Peking University.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609559

However, the information expressed in news articles is not always of high certainty, and some information in news articles is unreliable or uncertain. For example, the following sentence expresses the writer's speculation about an event, and it contains a weasel word "*seems*". Therefore, the sentence contain uncertain information.

"However, it *seems* that Obama will not use the platform to relaunch his stalled drive for Israeli-Palestinian peace."

Based on our analysis in Section 6.2, there are a considerable portion of sentences containing uncertain information in news articles, while human labeled reference summaries usually contain more certain information. Human annotators tend to extract and use certain sentences when they annotate summaries for news documents.

Disappointingly, existing summarization systems do not consider the factor of information certainty, and therefore some uncertain information may be injected into the summaries produced by existing summarization systems, which will hinder users' acquisition of the genuine information about the news event.

In this study, we investigate the factor of information certainty in the task of multi-document summarization, and propose a novel system called CTSUM to incorporate this new factor into the multi-document summarization task. The CTSUM system can produce more certain summaries for news articles. To the best of our knowledge, our proposed system is the first to incorporate the factor of information certainty into the summarization task.

After an in-depth investigation of the certain or uncertain information in news articles, we propose a few useful features to distinguish between certain sentences and uncertain sentences in news articles. Our proposed CTSUM system first estimates the certainty scores of sentences in news articles by using the SVM regression learner and then incorporates the estimated sentence-level certainty scores into a summarization system based on the graph-based ranking algorithm.

Experimental results verify the effectiveness of both the sentence-level certainty score estimation technique and the summarization system. Based on experiments on a manually-labeled sentence set, the estimated certainty scores are highly correlated with human labeled scores. Evaluations on the DUC2007 dataset shows that our proposed CTSUM system can extract both high-quality and certain summaries for news articles, and it can significantly outperform the baseline GRSUM system without considering the certainty factor.

The contributions of this work are summarized as follows:

1) We originally investigate the information certainty factor in the task of news document summarization.

2) We propose a novel system called CTSUM based on the graph-based ranking algorithm to produce high-quality and certain summaries for news articles.
3) We propose to assess the sentence-level information certainty in news articles by using a regression method, and evaluation results based on a manually-labeled data set verify the effectiveness of the method.
4) Experimental results on the DUC2007 dataset verify the effectiveness of our proposed CTSUM system for the topic-focused document summarization task.

The rest of this paper is organized as follows: Section 2 reviews related work. Our proposed system is overviewed in Section 3. Section 4 describes the key techniques of sentence-level certainty assessment and the evaluation results. Section 5 describes the summarization methods. The summarization evaluation results are presented in Section 6. Lastly, we conclude this paper in Section 7.

2. RELATED WORK

2.1 Multi-Document Summarization

Multi-document summarization is the task of producing a concise and fluent summary to deliver the major information for a given document set (usually news articles). If the summary is required to be biased to a given query or topic description, the task is called topic-focused or query-biased multi-document summarization. The methods for multi-document summarization can be coarsely categorized into abstraction-based or extraction-based. In this study, we focus on extraction-based methods, which have been adopted in most existing summarization systems.

The extraction-based summarization methods usually rank and select a few existing sentences in the documents to form a summary. Sentence extraction methods can be rule-based or learning-based. Rule-based methods make use of heuristic rules to score sentences by considering a few features, such as sentence position, TFIDF, etc. Such summarization methods include the centroid-based method [27] and NeATS [19]. Learning based methods have been proposed for optimally combining various sentence features based on supervised learning techniques [25, 33, 35, 51]. In order to capture more reliable semantic relatedness between sentences for document summarization, latent semantic analysis, matrix factorization and deep learning have been explored [21, 45]. Recent advanced methods formulate the summarization problem as various optimization problems, and solve the problems by selecting an optimal subset of sentences from the document set. Such methods include budgeted median [37], minimum dominating set [34], A* search [1], integer linear programming [10, 15], data reconstruction [11], submodular function [17, 18].

Particularly, graph-based ranking algorithms have been widely used for both generic and topic-focused document summarizations in recent years. LexRank [6] and TextRank [24] are two earliest graph-based summarization models, which adopt the basic PageRank style algorithms to rank sentences. Later on, a few models have been proposed to enhance the basic PageRank algorithm. For example, the topic-sensitive PageRank model is applied for topic-focused multi-document summarization [41]. The ClusterCMRW and ClusterHITS models [43] use cluster-based ranking algorithms to compute the saliency scores of sentences by taking into account the cluster-level information in the graph-based ranking algorithm. The DsR model [48] is a document-sensitive graph-based ranking model for multi-document summarization, and it considers the document-level influences in the ranking model. The mutual reinforcement chain model [47] further makes use of three different text granularities, i.e., document, sentences and terms, to construct a heterogeneous graph and develops a mutual reinforcement learning approach for topic-focused document summarization. The manifold-ranking and multi-modality manifold-ranking models have also been applied for topic-focused multi-document summarization [42, 44], and the basic assumption underlying the algorithms is that similar sentences are likely have same ranking scores and sentences on the same structure (i.e. cluster) are likely to have the same ranking scores.

All existing summarization systems do not consider the factor of information certainty in the summarization process, and they assume that all sentences in news articles are of equal certainty and thus any sentence can be selected into the summary if the sentence is highly ranked based on some evaluation metric, which is not very appropriate. In contrast, our proposed CTSUM system will take into account the sentence-level certainty score in the graph-based ranking algorithm in order to improve the summarization performance.

2.2 Related Studies on Information Certainty

The concept of certainty has different dictionary definitions, but they usually revolve around the notion of "the quality or state of mind of being free from doubt, especially on the basis of evidence" [23]. There are several related concepts, which have been addressed in previous NLP and linguistics studies: subjectivity, modality, evidentiality, factuality and hedging. Certainty can be viewed as a type of subjective information available in texts and a form of epistemic modality expressed through explicitly-coded linguistic means [31]. There are usually explicit certainty markers in texts to explicitly signal presence of certainty information that covers a full continuum of writer's confidence, ranging from uncertain possibility and withholding full commitment to statements to a confident necessity, reassurance, and emphasizing of the full commitment to statements. The certainty markers include such devices as subjectivity expressions, epistemic comments, evidentials, reporting verbs, attitudinal adverbials, hedges, shields, approximators, understatements, tentatives, intensifiers, emphatics, boosters, and assertives. In the NLP field, text subjectivity analysis, event factuality annotation and hedge detection are the most closely related studies.

Subjectivity concerns discourse participants and their stance with respect to what is conveyed by means of the text. Subjectivity manifests itself along different parameters, and hence encompasses a diverse set of interrelated research lines in the fields of NLP and data mining. For example, some work is devoted to identifying the author's affectual (or emotional) state [5]. Another related area focuses on opinion identification at different levels of granularity: document-level [26, 38], clausal- and phrasal-level [29, 50] and lexical level [30, 49].

Event factuality is defined as the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in discourse [32]. Events in discourse are usually passing through a spectrum of degrees of certainty. FactBank [32] is a corpus of events annotated with factuality information. Each event X in texts is annotated and assigned with one of six committed values or two underspecified values. The three positive committed values include CT+, PR+, PS+: CT+ means according to the source, it is certainly the case that X; PR+ means according to the source, it is probably the case that X; PS+ means according to the source, it is possibly the case that X. For example, the event "leave" in the following sentence was annotated with a committed value of "PS+".

It is possible that soviets in Kuwait will leave. (PS+)

Hedge detection has been a shared task of the CoNLL conference, which aims to identify sentences which contain uncertain information and recognizing in-sentence spans which are speculative [8]. It has received considerable interest recently in the NLP field. A hand-crafted list of hedge cues has been used to identify speculative sentence in MEDLINE abstracts [16]. The most recent approaches to this task exploit supervised or semi-supervised models and various features have been attempted, including single word features [22], n-gram features [36], syntactic features [13] and Wikipedia weasel tags [9]. Most researches are based on the BioScope corpus [40], which consists of manually labeled biological texts from full papers and scientific abstracts.

Another related but different research area is information credibility, which is a much broader concept than information certainty. Information credibility has been an active research area with the advent of on-line documents and social media content (e.g. blogs, comments, microblogs, etc.). Information credibility usually refers to the believability or quality of the information and/or its source. Hilligoss and Rieh [12] present a unifying framework of credibility assessment in which credibility is characterized across a variety of media and resources with respect to diverse information seeking goals and tasks. Automatic methods for credibility analysis have been performed on web pages, articles, blogs, messages and facts. Bendersky et al. [2] determine the quality of a web document by its readability, layout and ease-of-navigation, and other factors, and then a quality-biased ranking method is presented to promote documents containing high-quality content. Li et al [14] propose a two-step method to determine whether a given statement is truthful, and if it is not, find out the truthful statement most related to the given statement. Weerkamp and Rijke [46] estimate two groups of credibility indicators for blog posts, and integrate them into the topical blog post retrieval process. Castillo et al. [3] analyze microblog postings related to "trending" topics, and classify them as credible or not credible, based on a variety of features extracted from users' posting and re-posting behavior, from the text of the posts, and from citations to external sources.

3. SYSTEM OVERVIEW

As mentioned earlier, existing summarization systems do not consider the information certainty factor in the summarization process. In order to address this problem, we propose a novel system called CTSUM to produce more certain summaries for news articles. A certain summary must meet the following two goals:

1) **Summary content quality**: The content quality of a summary is the basic requirement in the document summarization task. The produced summary should overlap with the reference summaries as much as possible.

2) **Summary certainty**: The sentences in a certain summary should be of high certainty.

Since sentence is the natural and widely-used unit in the extraction-based summarization systems, we focus on sentence-level certainty analysis in this study.

Our proposed CTSUM system can achieve the above two goals by taking into account the sentence-level certainty score in the summarization process. It consists of two components: sentence-level certainty assessment and certain summary extraction. The first component aims to estimate the certainty score of each sentence in news articles and the second component aims to extract summary sentences by considering the sentence-level certainty scores. The two components will be described in details in next two sections, respectively.

4. SENTENCE-LEVEL CERTAINTY ASSESSMENT

4.1 Problem Definition and Corpus

A news writer's certainty level may remain constant in a news text or it may fluctuate from statement to statement. In this study, we focus on automatic estimation of sentence-level certainty level in news articles and the task is defined as a process of assessing how certain readers are about the statement in a sentence is evident and factual. In this study, we aim to assign a certainty score of each sentence to indicate the sentence's certainty level, rather than coarsely categorize each sentence into "certain" or "uncertain". The task of certainty assessment has traditionally been considered a task for humans. Fortunately, much certainty information comes from linguistic coding in texts and some explicit markers can be recognized and used for automatic certainty assessment.

As far as we know, there exist no publicly available benchmark corpus for this task. Related corpora include FactBank [32] and BioScope [40], but they are not suitable for this task. In the FactBank corpus, only the event-level factuality value is annotated, but the sentence-level certainty is different from the event-level factuality. In the BioScope corpus, the text genres are biological texts from full papers and scientific abstracts, which are different from new articles. Moreover, the sentences in the BioScope corpus are coarsely labeled as "certain" or "uncertain".

Therefore, we annotated our own corpus for this task. We first collected 1000 sentences from the FactBank corpus, which are from news articles. Two students were employed for certainty level annotation. The two students were firstly asked to read a short annotation guideline, which synthetically considered the categorization dimensions in [31]. Then they were asked to label a score between 1 and 5 for each sentence separately after they carefully read the sentence. Here, "5" means "very certain"; "4" means "almost certain"; "3" means "moderately certain"; "2" means "almost uncertain", "1" means "very uncertain". Note that during the annotation process, the original event-level factuality annotation results were provided to the students for their reference. The raw agreement between their annotations is 0.586, and the annotations' consistent degree measured by Correlation Coefficient (ρ) is 0.7424. Finally, the overall certainty score of each sentence is the average of the scores provided by the two annotators. Figure 1 shows the distribution of overall certainty scores of the sentences, and we can see there are a considerable portion of sentences with uncertain information.

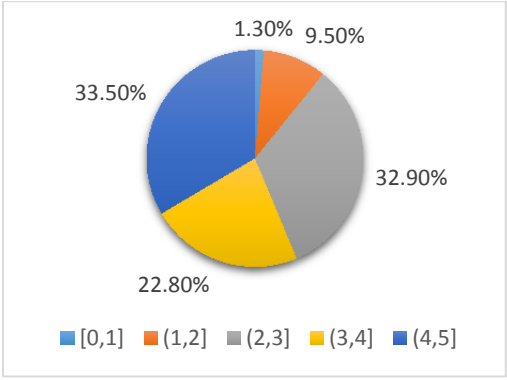

Figure 1. Distribution of overall certainty scores of sentences in our annotated corpus

4.2 Method

As mentioned above, sentence-level certainty assessment is a task of mapping each sentence to a numerical value corresponding to the certainty level. The larger the value is, the more certain the sentence is. The task can be considered a regression problem and in this study, we adopt the ε-support vector regression (ε-SVR) method [39] for addressing this prediction task. The SVR algorithm is firmly grounded in the framework of statistical learning theory (VC theory). The goal of a regression algorithm is to fit a flat function to the given training data points. More specifically, we use the LIBSVM tool [4] with the RBF kernel for this regression task.

We use the following four groups of features for each sentence, which are derived from different dimensions.

Explicit certainty markers: Explicit certainty marker words and phrases in a sentence are usually good indicators for its certainty level. For example, if there are some words like "probably" and "maybe" in a sentence, the sentence is likely to be uncertain, or the certainty level will be reduced accordingly. We collect a list of 34 explicit certainty markers (e.g. "likely", "possible"), and use the presence or absence of the markers as a binary feature.

Subjectivity markers: Usually, expressing objectively is more factual than expressing subjectively. Such information as judgments, opinions, attitudes, beliefs and emotions usually reflect an idea that does not represent an external reality, but rather a hypothesized world, existing only in someone's mind. Therefore, the use of such subjective information will devalue the certainty level. We collect 88 subjectivity markers (e.g. "fear") and use the presence or absence of the markers as a binary feature.

Time factor: It refers to the relevance of time (past, present and future) to the moment when the sentence was written. The future tense is less certain than the past and present tenses, since the future is usually predictions, plans, warnings and suggested actions, which may not come true. We collect 51 markers referring to future time (e.g. ""shall", "next year") and use the presence or absence of the markers as a binary feature.

Perspective factor: It refers to the view of writer or reporters. Directly involved state is more certain than indirectly one. For example, the information conveyed by indirectly involved third parties (e.g. experts, analysts, or anonymous "someone") is usually less certain than the information conveyed by directly involved ones (e.g. victims and witnesses). We collect 19 markers referring to indirectly involved states (e.g. "someone", "quoted") and use the presence or absence of the markers as a binary feature.

We use the open-source OpenNLP toolkit[1] to parse each sentence into a constituency-based parse tree and we obtain the ranges of main clause and subordinate clauses if existed. We compute the above feature values from the main clause and subordinate clauses separately, and use all the feature values for regression. The reason is the above factors have different influences in main clause or subordinate clauses.

All the above feature values are scaled by using the provided svm-scale program.

4.3 Evaluation Results

For evaluation, we randomly separated the labeled sentence set into ten sets of 100 sentences, and selected nine of them as a training set and the remaining one as a test set. We then used the LIBSVM tool for training and testing. The process was conducted for 10 times, and finally the results were averaged.

Two standard metrics were used for evaluating the prediction results. The two metrics are as follows:

Mean Square Error (MSE): This metric is a measure of how correct each of the prediction values is on average, penalizing more severe errors more heavily.

Pearson's Correlation Coefficient (ρ): This metric is a measure of whether the trends of prediction values matched the trends for human-labeled data.

Table 1 shows the prediction results with the standard deviation values. We implement a SVR baseline for comparison, which simply uses all words in sentences as features. The results of the SVR method after removing every group of features are also reported.

We can see that the overall results of the SVR method with all groups of features are very promising and the performance is much better than the baseline method. We can also see the each feature group is beneficial to the overall prediction performance, because the performance values have a decline after removing any group of features.

Table 1. Prediction results

Method	MSE	ρ
All feature groups	**0.3122** (stdev=0.0081)	**0.8769** (stdev=0.0296)
minus explicit certainty markers	0.4854 (stdev=0.0097)	0.7949 (stdev=0.0554)
minus subjectivity markers	0.5440 (stdev=0.0087)	0.7763 (stdev=0.0409)
minus time factor	0.6868 (stdev=0.0111)	0.6924 (stdev=0.0624)
minus perspective factor	0.4600 (stdev=0.0141)	0.8160 (stdev=0.0555)
Baseline (all words)	1.3038 (stdev=0.0193)	0.1634 (stdev=0.2424)

Finally, we apply the SVR method to predict the certainty score of each sentence in the document sets to be summarized. The certainty score is then normalized into [0, 1] by dividing by the maximum score. Finally, each sentence s_i in news articles is associated with a normalized certainty score $CertainScore(s_i)$. The larger the score is, the more certain the sentence is.

5. CERTAIN SUMMARY EXTRACTION

In this study, we focus on topic-focused multi-document summarization and adopt the graph-based ranking framework for sentence ranking because it has been widely adopted for document summarization in recent years [41]. Existing graph-based ranking algorithms are based on "voting" or "recommendation" between sentences. A link between two sentences is considered as a vote cast from one sentence to the other sentence. The score associated with a sentence is determined by the votes that are cast for it, and the score of the sentences casting these votes. For topic-focused multi-document summarization, the relevance between sentences and the given topic is incorporated into the graph-based ranking framework as priors.

In order to make use of the certainty information of each sentence, we assume that sentences with high certainty level should be ranked higher than sentences with low certainty level. We alter the transition matrix by considering the sentence-level certainty information for achieving this goal.

[1] http://opennlp.apache.org/

Formally, given a document set D and a topic or query description q, let $G=(V, E)$ be a directed graph to reflect the relationships between sentences in the document set. V is the set of vertices and each vertex s_i in V is a sentence in the document set. E is the set of edges. Each edge e_{ij} in E is associated with an affinity weight $f(s_i, s_j)$ from sentences s_i to s_j ($i \neq j$). The weight is computed using the standard cosine measure between the two sentences as follows:

$$f(s_i, s_j) = sim_{cosine}(s_i, s_j) = \frac{\vec{s}_i \cdot \vec{s}_j}{|\vec{s}_i| \times |\vec{s}_j|} \quad (1)$$

where $\vec{s}_i$ and $\vec{s}_j$ are the corresponding term vectors of s_i and s_j. Here, we have $f(s_i, s_j)=f(s_j, s_i)$ and let $f(s_i, s_i)=0$ in order to avoid self transition.

The transition probability from s_i to s_j is then defined by normalizing the corresponding affinity weight as follows:

$$p(s_i, s_j) = \begin{cases} \dfrac{f(s_i, s_j)}{\sum_{k=1}^{|V|} f(s_i, s_k)}, & if \sum f \neq 0 \\ 0, & otherwise \end{cases} \quad (2)$$

Note that $p(s_i, s_j)$ is usually not equal to $p(s_j, s_i)$. We use the row-normalized matrix $\tilde{M}=(\tilde{M}_{i,j})_{|V|\times|V|}$ to describe G with each entry corresponding to the transition probability.

$$\tilde{M}_{i,j} = p(s_i, s_j) \quad (3)$$

We can see that the transition probability from s_i to s_j relies solely on the relative similarity between s_i and s_j, compared with the similarity between s_i and all other sentences. If sentence s_j is highly similar to s_i, but all other sentences are less similar to s_i, then the transition probability from s_i to s_j is high, no matter if s_j is certain or not. A high transition probability from s_i to s_j means s_i will propagate more of its score to s_j.

In order to make use of the estimated certainty score $CertainScore(s_j)$ of each sentence s_j, we change Equation (1) into Equation (4) by adding the certainty factor to obtain a new affinity weight $f^{new}(s_i, s_j)$ from sentence s_i to sentence s_j ($i \neq j$) as follows:

$$f^{new}(s_i, s_j) = f(s_i, s_j) \times (1 + \lambda \cdot CertainScore(s_j)) = sim_{cosine}(s_i, s_j) \times (1 + \lambda \cdot CertainScore(s_j)) \quad (4)$$

where $\lambda \geq 0$ is a parameter to control the influence of the certainty score of each sentence and $f^{new}(s_i, s_j)$ is usually not equal to $f^{new}(s_j, s_i)$. The new transition probability from s_i to s_j is then computed as follows:

$$p^{new}(s_i, s_j) = \begin{cases} \dfrac{f^{new}(s_i, s_j)}{\sum_{k=1}^{|V|} f^{new}(s_i, s_k)}, & if \sum f^{new} \neq 0 \\ 0, & otherwise \end{cases} \quad (5)$$

We use the new row-normalized matrix $\tilde{M}^{new}=(\tilde{M}^{new}_{i,j})_{|V|\times|V|}$ to describe G with each entry corresponding to the new transition probability.

$$\tilde{M}^{new}_{i,j} = p^{new}(s_i, s_j) \quad (6)$$

We can see that if λ is set to 0, then we have $f^{new}(s_i, s_j)= f(s_i, s_j)$, $p^{new}(s_i, s_j)= p(s_i, s_j)$ and $\tilde{M}^{new} = \tilde{M}$. When λ is set to a very large value, the certainty factor in Equation (4) is dominated by the certainty score of sentence s_j.

The new transition probability from s_i to s_j is relying not only on the relative similarity between s_i and s_j, but also on the certainty level of s_j. Sentences with high certainty level are likely to receive more from a source sentence than sentences with low certainty level, and thus highly certain sentences are likely to be ranked higher than less certain sentences for summary extraction.

Note that in Equations (4) and (5), we do not make use of the certainty score of the source sentence s_i when computing the transition probability from s_i to s_j, because the certainty score of s_i is the same for all the target sentences, and the use of the score will not affect the transition probability after row normalization.

We also compute the relevance score $rel(s_i, q)$ of each sentence s_i to query q by using the standard cosine measure.

$$rel(s_i, q) = sim_{cosine}(s_i, q) = \frac{\vec{s}_i \cdot \vec{q}}{|\vec{s}_i| \times |\vec{q}|} \quad (7)$$

The relevance score is then normalized to $rel'(s_i, q)$ as follows in order to make the sum of all relevance values of the sentences equal to 1.

$$rel'(s_i, q) = \frac{rel(s_i, q)}{\sum_{k=1}^{|V|} rel(s_k, q)} \quad (8)$$

Based on matrix $\tilde{M}$, the topic-biased saliency score $InfoScore(s_i)$

$$InfoScore(s_i) = \mu \cdot \sum_{all\ j \neq i} InfoScore(s_j) \cdot \tilde{M}_{j,i} + (1 - \mu) \cdot rel'(s_i, q) \quad (9)$$

for sentence s_i can be deduced from those of all other sentences linked with it, and it can be formulated in a recursive form as follows:

$$InfoScore^{new}(s_i) = \mu \cdot \sum_{all\ j \neq i} InfoScore^{new}(s_j) \cdot \tilde{M}^{new}_{j,i} + (1 - \mu) \cdot rel'(s_i, q) \quad (10)$$

Similarity, based on matrix $\tilde{M}^{new}$, the new topic-biased saliency score $InfoScore^{new}(s_i)$ for sentence s_i can be deduced recursively as follows:

where μ is the damping factor usually set to 0.85, as in the PageRank algorithm.

After computing the saliency score $InfoScore(s_i)$ or $InfoScore^{new}(s_i)$, we can select highly ranked sentences to form the summary.

For the multi-document summarization tasks, some sentences are highly overlapping with each other, and thus we apply the same greedy algorithm as in [41, 44] to penalize the sentences highly overlapping with other highly scored sentences. The sentences are firstly ranked by their saliency scores, and then the most highly ranked sentence is selected into the summary, and the saliency scores of the remaining sentences are penalized according to the content overlap (standard cosine similarity) between the sentences and the selected summary sentence. The above selection is iterated until the summary length reaches the length limit. The details of the algorithm is omitted here.

Based on the saliency scores computed with Equation (9), we can produce summaries without considering the certainty information of sentences, and the system is named **GRSUM** in this study.

Based on the saliency scores computed with Equation (10), we can produce summaries with considering the certainty information of sentences, which is named **CTSUM** in this study.

The GRSUM is considered the baseline system, and it is actually a degeneration version of CTSUM when is λ set to 0.

It is worth noting that the certainty level of sentences can also be easily incorporated into other document summarization methods, which is, however, not the focus of this study.

6. EMPIRICAL EVALUATIN

6.1 Data Set

We conducted experiments for topic-focused multi-document summarization, which has been one of the fundamental tasks in the DUC conferences.

We used the DUC2007 dataset as test set for evaluation. Forty-five document clusters with topic descriptions were provided. NIST assessors first developed topics/questions of interest to them and then chose a set of documents relevant to each topic. Reference summaries have been created for all the document clusters by NIST assessors. Note that multiple reference summaries written by different assessors are provided for each document cluster. Given a topic description and relevant documents, the task aims to create from the documents a brief, well-organized, fluent summary which answers the need for information expressed in the topic. Each topic consisted of a title and a narrative text, and we concatenated the title and narrative text to represent the topic. In addition, we used the DUC2006 dataset as development set for parameter tuning and the value of λ is set to 1 for CTSUM in the experiments. The two datasets are summarized in Table 2.

As a preprocessing step for similarity computation, the stop words in each sentence were removed and the remaining words were stemmed using the Porter's stemmer[2].

Table 2. Summary of datasets

	Development set (DUC 2006)	Test set (DUC 2007)
Task	Topic-focused	Topic-focused
Number of clusters	50	45
Average document number per cluster	25	25
Data source	AQUAINT	AQUAINT
Summary length	250 words	250 words

6.2 Data Certainty Analysis

In this section we conduct analysis of the certainty level of the evaluation datasets. We used the SVR method to predict the certainty score of each sentence in each document set and obtain an average score for each document set, and then the scores are further averaged across all document sets. We also used the SVR method to predict the certainty score of each sentence in each reference summary and obtain an average score for each reference summary, and then the scores are further averaged across all reference summaries for all document sets. The final average scores are presented in Table 3, where "stdev" means the standard deviation.

We can see that the average certainty score in reference summaries are significantly higher than that in documents on both the development set and the test set. That's to say, in the news articles, there are a considerable portion of sentences with relatively low certainty, but in reference summaries, most sentences are highly certain. The results validate our assumption that human annotators tend to select or write certain sentences to produce the summaries, and the uncertain information is seldom used. Therefore, our proposed CTSUM can benefit from the explicit use of the certainty scores of sentences.

Table 3. Comparison of average certainty scores of sentences in documents and reference summaries

	Development set (DUC 2006)	Test set (DUC 2007)
Average certainty score for Sentences in Documents	3.778 (stdev=0.324)	3.798 (stdev=0.337)
Average certainty score for Sentences in Reference Summaries	4.253 (stdev=0.251)	4.225 (stdev=0.274)

6.3 Evaluation Metrics

We used the ROUGE-1.5.5 [20] toolkit for evaluating the content quality of produced summaries by comparing them with the reference summaries, which has been widely adopted by DUC and TAC for automatic summary quality evaluation. It measured summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram based measure and the recall oriented score, the precision oriented score and the F-measure score for ROUGE-N are computed as follows:

$$ROUGE-N_{Recall} = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{n\text{-}gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{n\text{-}gram \in S} Count(n-gram)} \quad (11)$$

$$ROUGE-N_{Precision} = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{n\text{-}gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{Candidate\ Summary\}} \sum_{n\text{-}gram \in S} Count(n-gram)} \quad (12)$$

$$ROUGE-N_{F-Measure} = \frac{2 \times ROUGE-N_{Recall} \times ROUGE-N_{Precision}}{ROUGE-N_{Recall} + ROUGE-N_{Precision}} \quad (13)$$

where n stands for the length of the n-gram, and $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. $Count(n\text{-}gram)$ is the number of n-grams in the reference summaries or candidate summary.

We showed the popular three ROUGE scores in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based) and ROUGE-SU4 (based on skip bigram with a maximum skip distance of 4). Both F-measure and Recall scores are reported in the experiments. Note that the ROUGE scores are computed for each document set, and then the scores are averaged.

When using the ROUGE-1.5.5 toolkit, we used the option "-l 250" in order to truncate the summary longer than the length limit, and used the option "-m" for word stemming.

[2] http://www.tartarus.org/martin/PorterStemmer/

6.4 Evaluation Results

Firstly, we evaluate the content quality of our proposed CTSUM system by using the ROUGE metrics. Since our proposed CTSUM system is a direct improvement of the GRSUM system by considering a new factor of sentence-level certainty in the transition probability calculation, we compare CTSUM and GRSUM to show whether the new factor is helpful to the summarization performance.

The comparison results over ROUGE F-measure and Recall metrics on the test set are presented in Table 4. In the table, the NIST baseline is also presented and it is the official baseline system established by NIST. We can see from Table 4 that CTSUM outperforms GRSUM over all metrics. In order to show whether the performance differences between CTSUM and GRSUM are statistically significant, we apply the paired t-test (two-tailed) over each metric and the p-values are shown in Table 5. We can see that all the p-values are smaller than 0.01, and the results demonstrate that CTSUM can significantly outperform GRSUM over all ROUGE metrics. Since the difference between CTSUM and GRSUM lies only in the use of the sentence-level certainty score for calculating the transition probability between sentences, the superior performance of CTSUM over GRSUM can validate the efficacy of the sentence-level certainty factor.

Table 4. Comparison results on DUC2007

	ROUGE-1 F-measure	ROUGE-2 F-measure	ROUGE-SU4 F-measure
CTSUM	**0.42663**	**0.10825**	**0.16162**
GRSUM	0.41966	0.10261	0.15598
NIST Baseline	0.33434	0.06479	0.11264
	ROUGE-1 Recall	**ROUGE-2 Recall**	**ROUGE-SU4 Recall**
CTSUM	**0.43101**	**0.10931**	**0.16324**
GRSUM	0.42355	0.10350	0.15738
NIST Baseline	0.31177	0.06021	0.10481

Table 5. P-values of two-tailed t-tests for ROUGE scores of CTSUM and GRSUM on DUC2007

	ROUGE-1 F-measure	ROUGE-2 F-measure	ROUGE-SU4 F-measure
p-value	0.009781	0.006911	0.001121
	ROUGE-1 Recall	**ROUGE-2 Recall**	**ROUGE-SU4 Recall**
p-value	0.006735	0.00569	0.000732

We also compare our proposed CTSUM system with a few advanced baselines, as shown in Table 6. For fair comparison, we only compare CTSUM with the following unsupervised summarization methods:

ManifoldRank: This method is proposed in [44], and it makes use of the manifold-ranking algorithm to rank sentences. The ROUGE F-measure scores in the table are directly borrowed from [42].

MultiMR: This method is proposed in [42], and it considers the within-document sentence relationships and the cross-document sentence relationships as two modalities and makes use of the multi-modality manifold-ranking algorithm to rank sentences. Four different ranking schemes (i.e. LIN, COM, SEQ1 and SEQ2) are employed. The ROUGE F-measure scores in the table are directly borrowed from [42].

DSDR: This method is proposed in [11], and it extracts summary sentences that can best reconstruct the original documents. The linear reconstruction model (DSDR-lin) and the nonnegative linear reconstruction model (DSDR-non) are proposed. The ROUGE F-measure scores in the table are directly borrowed from [11]

ClusterHITS: This method is proposed in [43], and it considers the topic clusters as hubs and sentences as authorities, then applies the HITS algorithm to rank sentences. The ROUGE F-measure scores in the table are directly borrowed from [11].

SNMF: This method is proposed in [45], and it uses symmetric non-negative matrix factorization for sentence clustering and then select sentences from each cluster. The ROUGE F-measure scores in the table are directly borrowed from [11].

The ROUGE F-measure scores of the systems are shown in Table 6. We can see that our proposed CTSUM system outperform the baseline methods and the performance of CTSUM is comparable to that of the state-of-the-art methods.

Table 6. Comparison with other methods on DUC2007

	ROUGE-1 F-measure	ROUGE-2 F-measure
CTSUM	0.42663	0.10825
ManifoldRank	0.41303	0.10009
MultiMR(LIN)	0.42041	0.10302
MultiMR(COM)	0.41837	0.10263
MultiMR(SEQ1)	0.41803	0.10292
MultiMR(SEQ2)	0.41600	0.10095
DSDR-lin	0.36055	0.07163
DSDR-non	0.39573	0.07439
ClusterHITS	0.32873	0.06625
SNMF	0.28651	0.04232

Secondly, we evaluate the summary certainty of our proposed CTSUM system. Two evaluation methods are employed: automatic evaluation and manual evaluation.

Automatic evaluation makes use of the automatically estimated certainty scores. Each summary's certainty score is the average of the certainty scores of all the sentences in the summary, and then the summaries' certainty scores are averaged across the 45 document sets. The average certainty scores for CTSUM and GRSUM are compared in Table 7, and the detailed certainty scores for all summaries are compared in Figure 2. We can see that the average certainty score for CTSUM is significantly higher than that for GRSUM, which means CTSUM can produce more certain summaries than GRSUM.

Manual evaluation relies on subjective assessment of the summaries' certainty levels. Two students are employed to manually check all the summaries produced by CTSUM and GRSUM, and label for each summary a score between 1 to 5 to indicate the summary's certainty level. We average the labeled scores for each summary across the two students, and then average the scores across the 45 document sets. The comparison results are shown in Table 8 and Figure 3. We can get the same conclusion that CTSUM can indeed produce more certain summaries than GRSUM.

Table 7. Automatic evaluation of summary certainty on DUC2007

	Average Certainty Score
GRSUM	3.71
CTSUM	4.20
p-value for two-tailed t-test	3.41668E-20

Table 8. Manual evaluation of summary certainty on DUC2007

	Average Certainty Score
GRSUM	3.24
CTSUM	4.06
p-value for two-tailed t-test	5.7854E-11

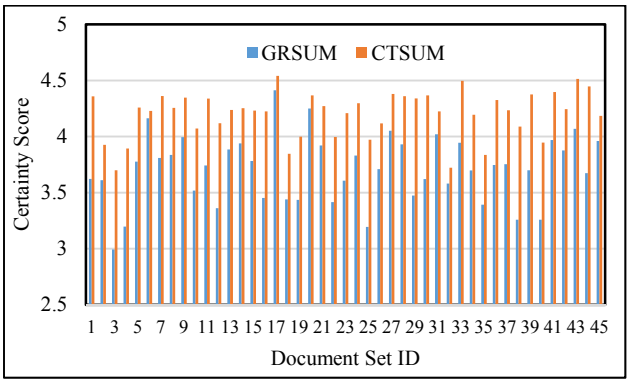

Figure 2. Detailed comparison of estimated certainty scores of summaries on DUC2007

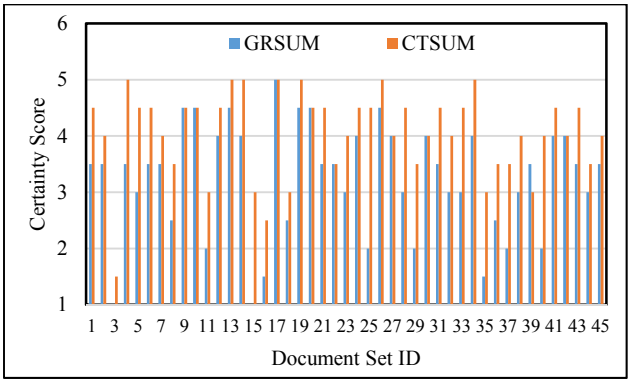

Figure 3. Detailed comparison of manually labeled certainty scores of summaries on DUC2007

We now examine the influence of parameter λ in our proposed CTSUM system. In the above experiments, λ is set to 1, and we now vary λ from 0 to 5 with a step of 0.5. We show the curves of ROUGE-1 F-measure, ROUGE-2 F-measure, ROUGE-SU4 F-measure, and the average certainty score in Figures 4-7, respectively. Note that when λ is equal to 0, the CTSUM system corresponds to the GRSUM system. We can see from Figures 4-6 that when λ is equal to 0, the summarization performance is the worst, and the summarization performance rise up with the increase of λ from 0 to 1. When λ is larger than 1, the performance almost keeps steady. The reason is that when λ is set to any large value, the certainty factor in Equation (4) is dominated by the certainty score of s_j, and the certainty factor will not be affected by the specific value of λ. We can see from Figure 7 that when λ is equal to 0, the average certainty score of produced summaries is the lowest, and the average certainty score rises up sharply with the increase of λ from 0 to 1. When λ is larger than 1, the average certainty score tends to become steady. The curves in the figures demonstrate that our CTSUM system can always produce summaries with better content quality and higher certainty than GRSUM, and the certainty factor is validated to be very helpful in the summarization process.

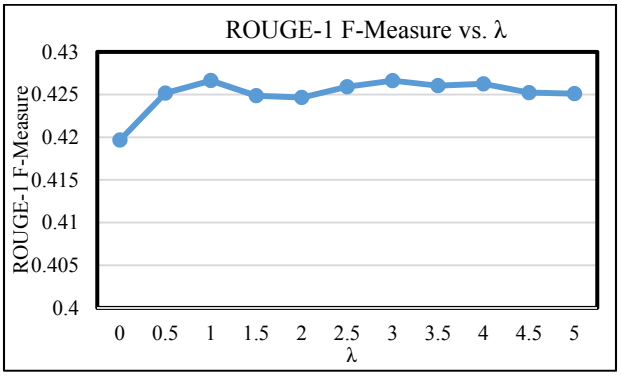

Figure 4. ROUGE-1 F-measure vs. λ on DUC2007

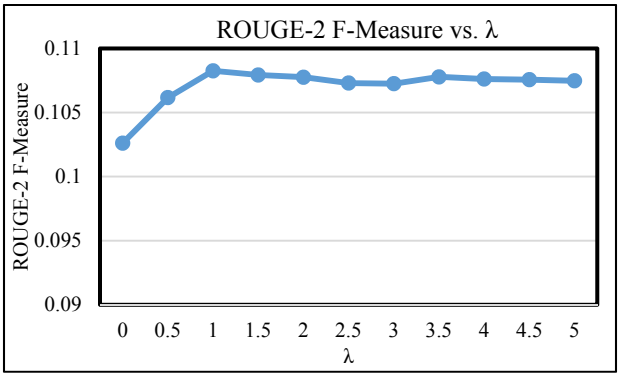

Figure 5. ROUGE-2 F-measure vs. λ on DUC2007

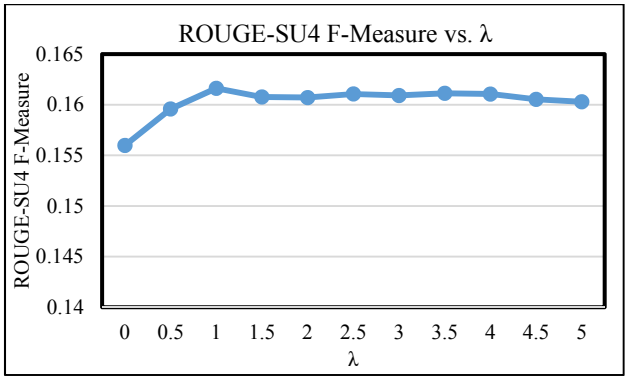

Figure 6. ROUGE-SU4 F-measure vs. λ on DUC2007

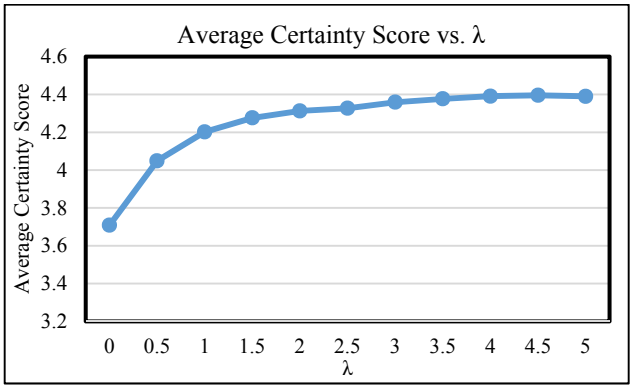

Figure 7. Average certainty score vs. λ on DUC2007

Finally, we show the example summaries produced by GRSUM and CTSUM for document set D0725 in DUC2007 (Corresponding to document set ID 25 in Figures 2-3). The unique sentences in each summary are highlighted in italics and bold. The estimated certainty score of the summary produced by GRSUM is 3.195, while the estimated certainty score of the summary produced by CTSUM is 3.971. We can see that some unique sentences extracted by GRSUM have low certainty scores, e.g., the 2nd sentence and the 11th sentence.

Summary sentences extracted by GRSUM (without sentence ordering):

Iran's new missiles are not yet operational and probably more, years away from developing nuclear weapons.

The United States and Israel, which suspect that Iran may want to acquire nuclear weapons, have opposed the Russian-Iranian nuclear cooperation.

Iran will not build an atomic bomb despite nuclear tests by Pakistan and India that have forced ``a new security situation" on the country.

Iran has long been accused by the U.S. of trying to gain nuclear capability.

But the tests do not put Iran under pressure to start its own nuclear weapons program.

Russia's assistance for Iran in the field has triggered accusations from the United States and Israel that Iran is seeking the technology for developing nuclear weapons.

Israel has long feared Iran's growing weapons program and warned of ties between Iran and Pakistan after last months nuclear testing by Pakistan and India.

Russia said Iran's technological potential does not allow it to produce nuclear weapons.

He said that Pakistan did not know the status of Iran's nuclear program.`` We have no idea about other countries' nuclear programs.

Senior Iranian officials on Monday assured Russia that Iran will use nuclear energy for peaceful purposes and will pose no harm to any countries.

It means that in a few years time, Iran will be able to threaten with a nuclear weapon.

So far, Iran is only known to have several small nuclear research facilities.

Defense Minister Yitzhak Mordechai said there was no call for such declarations.

Summary sentences extracted by CTSUM (without sentence ordering):

Israel has long feared Iran's growing weapons program and warned of ties between Iran and Pakistan after last months nuclear testing by Pakistan and India.

Iran's new missiles are not yet operational and probably more, years away from developing nuclear weapons.

So far, Iran is only known to have several small nuclear research facilities.

But the tests do not put Iran under pressure to start its own nuclear weapons program.

Russia's assistance for Iran in the field has triggered accusations from the United States and Israel that Iran is seeking the technology for developing nuclear weapons.

The Clinton administration expressed new worry Monday over Iran's nuclear program and whether that country has acquired the capability to make nuclear bombs and other weapons of mass destruction.

Iran on Wednesday strongly lambasted the United States for its accusations against Iranian-Russian nuclear cooperation.

Russia is helping Iran build a 1,000-megawatt nuclear power plant in Bushehr, southern Iran.

Iran will not build an atomic bomb despite nuclear tests by Pakistan and India that have forced ``a new security situation" on the country.

Iran has long been accused by the U.S. of trying to gain nuclear capability.

Netanyahu voiced Israel's concern over Iran's nuclear programs after Tehran and Moscow signed a memorandum of understanding last Tuesday to promote their nuclear cooperation.

Mohammadi said that Iran and Russia have been committed to controlling the export of the nuclear energy from either country in accordance with the Nuclear Non-Proliferation Treaty (NPT).

7. CONCLUSIONS AND FUTURE WORK

In this paper we investigate the certainty factor in the summarization process and propose a new system called CTSUM to extract more certain summaries for news articles. Our CTSUM system first estimates the certainty score of each sentence, and then makes use of the sentence-level certainty score in the graph-based ranking summarization algorithm. Experimental results on the DUC2007 dataset verify the helpfulness of the sentence-level certainty score, and our proposed CTSUM system can significantly outperform the baseline GRSUM system.

As mentioned earlier, the sentence-level certainty score can be easily incorporated into other summarization methods, and in future work we will conduct more experiments with other summarization methods to further show the merit of the certainty factor.

In this study, we focus on summarization of news articles. With the increase of social media content (e.g. Twitter, blogs), the certainty or credibility problem in social media content is more serious than that in news articles, and we will investigate incorporating the certainty factor into summarization of social media content in future work.

8. ACKNOWLEDGMENTS

This work was supported by NSFC (61170166, 61331011), Beijing Nova Program (2008B03) and National High-Tech R&D Program (2012AA011101). We thank Jianmin Zhang and Bingqing Li for data annotation, and Qi Su for earlier discussion of data annotation guideline. We also thank the anonymous reviewers for their helpful comments.

9. REFERENCES

[1] A. Aker, T. Cohn, and R. Gaizauskas. Multi-document summarization using A* search and discriminative training. In *Proceedings of EMNLP2010*.

[2] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of Web documents. In Proceedings of WSDM-2011.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *Proceedings of WWW2011*.

[4] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[5] K. Dave, S. Lawrence and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519-528. ACM, 2003.

[6] G. Erkan and D. R. Radev. LexPageRank: prestige in multi-document text summarization. In *Proceedings of EMNLP-04*.

[7] D. K. Evans, J. L. Klavans, and K. R. McKeown. Columbia Newsblaster: multilingual news summarization on the Web. In *Proceedings of HLT-NAACL-04 Demonstrations*.

[8] R. Farkas, V. Vincze, G. Móra, J. Csirik and G. Szarvas. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, pp. 1-12. Association for Computational Linguistics, 2010.

[9] V. Ganter and M. Strube. Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 173-176. Association for Computational Linguistics, 2009.

[10] D. Gillick, B. Favre and D. Hakkani-Tur. The ICSI summarization system at TAC 2008. In *Proceedings of the Text Understanding Conference*. 2008.

[11] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai and X. He. Document Summarization Based on Data Reconstruction. In *AAAI*. 2012.

[12] B. Hilligoss and S. Y. Rieh. Developing a unifying framework of credibility assessment: construct, heuristics, and interactions in context. *Information Processing and Management*, 44(4): 1467-1484, 2008.

[13] H. Kilicoglu and S. Bergler. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC bioinformatics* 9, no. Suppl 11 (2008): S10.

[14] X. Li, W. Meng, and C. Yu. T-verifier: verifying truthfulness of fact statements. In Proceedings of ICDE2011.

[15] C. Li, X. Qian and Y. Liu. Using supervised bigram-based ILP for extractive summarization. In *Proceedings of ACL*, pp. 1004-1013. 2013.

[16] M. Light, X. Y. Qiu and P. Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, pp. 17-24. 2004.

[17] H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 912-920. Association for Computational Linguistics, 2010.

[18] H. Lin and J. Bilmes. A Class of Submodular Functions for Document Summarization. In *ACL*, pp. 510-520. 2011.

[19] C.-Y. Lin and E.. H. Hovy. From single to multi-document summarization: a prototype system and its evaluation. In *Proceedings of ACL-02*.

[20] C.-Y. Lin and E.H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of HLT-NAACL -03.

[21] Y. Liu, S.-H. Zhong and W. Li. Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning. In *AAAI*. 2012.

[22] B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, pp. 992-999. 2007.

[23] Merriam-Webster Online Dictionary, http://www.m-w.com/ Accessed, January 30, 2004.

[24] R. Mihalcea and P. Tarau. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP-05*.

[25] Y. Ouyang, S. Li, W. Li. Developing learning strategies for topic-focused summarization. In *Proceedings of CIKM-07*.

[26] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79-86. Association for Computational Linguistics, 2002.

[27] D. R. Radev, H. Y. Jing, M. Stys and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40: 919-938, 2004.

[28] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. NewsInEssence: summarizing online news topics. *Communications of the ACM*, 48(10), 2005.

[29] J. Read and J. Carroll. Annotating expressions of appraisal in English. *Language Resources and Evaluation* 46, no. 3 (2012): 421-447.

[30] E. Riloff, J. Wiebe and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 25-32. Association for Computational Linguistics, 2003.

[31] V. L. Rubin, E. D. Liddy and N. Kando. Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*, pp. 61-76. Springer Netherlands, 2006.

[32] R. Saurí and J. Pustejovsky. FactBank: a corpus annotated with event factuality. *Language resources and evaluation* 43, no. 3 (2009): 227-268.

[33] F. Schilder and R. Kondadadi. FastSum: fast and accurate query-based multi-document summarization. In *Proceedings of ACL-08: HLT*.

[34] C. Shen and T. Li. Multi-document summarization via the minimum dominating set. In Proceedings of COLING-10.

[35] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *Proceedings of IJCAI-07*.

[36] G. Szarvas. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of 46th Meeting of the Association for Computational Linguistics*. 2008.

[37] H. Takamura and M. Okumura. Text summarization model based on the budgeted median problem. In *Proceedings of CIKM-09*.

[38] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424. Association for Computational Linguistics, 2002.

[39] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[40] V. Vincze, G. Szarvas, R. Farkas, G. Móra and J. Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics* 9, no. Suppl 11 (2008): S9.

[41] X. Wan. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval*, 11: 25-49, 2008.

[42] X. Wan and J. Xiao. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization. In *IJCAI*, pp. 1586-1591. 2009.

[43] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *Proceedings of SIGIR-08*.

[44] X. Wan, J. Yang and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI-07*.

[45] D. Wang, T. Li, S. Zhu, C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In Proceedings of SIGIR-08.

[46] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08:HLT*, pages 923-931.

[47] F. Wei, W. Li, Q. Lu and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In Proceedings of SIGIR-08.

[48] F. Wei, W. Li, Q. Lu, and Y. He. A document-sensitive graph model for multi-document summarization. *Knowledge and information systems* 22, no. 2 (2010): 245-259.

[49] J. Wiebe. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pp. 735-740. 2000.

[50] J. Wiebe, T. Wilson and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39, no. 2-3 (2005): 165-210.

[51] K.-F. Wong, M. Wu and W. Li. Extractive summarization using supervised and semisupervised learning. In *Proceedings of COLING-08*.