Re-ranking Summaries based on Cross-document Information Extraction

Heng Ji[1], Juan Liu[1], Benoit Favre[2], Dan Gillick[3], Dilek Hakkani-Tur[3]

[1] Computer Science Department, Queens College and Graduate Center,
City Univeristy of New York,
New York, NY 11367, USA

[2] LIUM, Université du Maine
Avenue Laënnec, 72085 Le Mans Cedex 9, France

[3] Computer Science Department and International Computer Science Institute,
University of California, Berkeley
Berkeley, CA 94704, USA

hengji@cs.qc.cuny.edu

Abstract. This paper describes a novel approach of improving multi-document summarization based on cross-document information extraction (IE). We describe a method to automatically incorporate IE results into sentence ranking. Experiments have shown our integration methods can significantly improve a high-performing multi-document summarization system, according to the ROUGE-2 and ROUGE-SU4 metrics (7.38%% relative improvement on ROUGE-2 recall), and the generated summaries are preferred by human subjects (0.78 higher TAC Content score and 0.11 higher Readability/Fluency score).

Keywords: Multi-document Summarization, Information Extraction

1 Introduction

Since about one decade ago Information Extraction (IE) and Automated Text Summarization have been recognized as two tasks sharing the same goal ([1]) – extract accurate information from unstructured texts according to a user's specific desire, and present the information to the user in a compact form. These two tasks have been studied separately and quite intensively over the past decade. Various corpora have been annotated for each task, a wide range of models and machine learning methods have been applied, and separate official evaluations have been organized. There has clearly been a great deal of progress on the performance of both tasks.

Because a significant percentage of queries in the summarization task involve facts (entities, relations and events), it is beneficial to exploit results extracted by IE techniques in automatic summarization. Some earlier work (e.g. [2], [3]) used Message Understanding Conference (MUC) ([4]) IE to generate or improve summaries. The IE task has progressed from MUC-style single template extraction to the more comprehensive Automatic Content Extraction (ACE) that targets at more fine-grained types of facts. The IE methods have also been advanced from single-document IE to cross-document dynamic event chain extraction (e.g. [5]) and static attribute extraction ([9]). In addition, a lot of current IE systems couple supervised learning techniques with traditional pattern matching approaches, which enable them to produce reliable confidence values (e.g. [5]). Therefore a summarization process can have more flexibility to choose using IE results or the original sentences ([6]). Based on the above reasons we feel the time is now ripe to explore some novel methods to marry these two tasks again and raise summarization to a higher level of performance.

From a collection of documents for a specific query, we extract facts in both queries and the documents. We use a high-performing multi-document extractive summarizer as our baseline, and tightly integrate IE results into its sentence ranking and compression. Experiment results show this integration method can achieve significant improvement on both standard summarization metrics and human judgement.

2 Task and Baseline System

2.1 TAC Summarization Task

The summarization task we are addressing is that of the NIST Text Analysis Conference (TAC) multi-document summarization evaluation ([7]). This task involves generating fixed-length summaries from 10 newswire documents, each on a given query including a specific topic. For example, given a query "Judge Joan Lefkow's Family Murdered/Describe the murders of Judge Joan Lefkow's husband and mother, and the subsequent investigation. Include details about any evidence, witnesses, suspects and motives." and 10 documents, a summarization system is required to generate a summary about specific entities ("Judge Joan Lefkow"), relations ("family") and events ("murder" and "investigation").

2.2 Baseline Summarization System

We apply a top-performing TAC summarization system ([8]) as our baseline. In this model, a summary is the set of sentences that best covers the relevant concepts in the document set, where concepts are simply word bigrams valued by their document frequency. The concepts with low-frequency or stop-words are filtered. The value of a sentence is the sum of the concept values it contains. The goal of summarization is

modeled in a way to find the collection with maximum value, subject to a length constraint. This problem is solved efficiently with an integer linear programming (ILP) solver. A sentence compression component is used to post-process candidate sentences. The compression step consists of dependency tree trimming using high-confidence semantic role labeling decisions. Non-mandatory temporal and manner arguments are removed and indirect discourse is reformulated in direct form.

3 Cross-document IE Annotation

We apply a state-of-the-art English cross-document IE system ([6], [9]) to extract facts from the input documents. This system was developed for the NIST Automatic Content Extraction Program (ACE 2005)[1] and TAC KBP 2010 Program[2].

ACE2005 defined 7 types of entities, 18 types of relations and 33 distinct types of relatively 'dynamic' events. KBP2010 defined 42 types of relatively 'static' slots (e.g. "*Ruth D. Masters is the wife of Hyman G. Rickover*" indicates that the "*per:spouse*" slot for person "*Hyman G. Rickover*" is "*Ruth D. Masters* ").

The IE pipeline includes name tagging, nominal mention tagging, coreference resolution, time expression extraction and normalization, relation extraction and event extraction. Names are identified and classified using an HMM-based name tagger. Nominals are identified using a maximum entropy-based chunker and then semantically classified using statistics from the ACE training corpora. Relation extraction and event extraction are also based on maximum entropy models, incorporating diverse lexical, syntactic, semantic and ontological knowledge. At the end an event coreference resolution component is applied to link coreferential events, based on a pairwise maximum entropy model with linguistic attributes and a graph-cut clustering model. Then an event tracking component is applied to identify important entities which are frequently involved in events as 'centroid entities'; link and order the events centered around each centroid entity on a time line.

Our slot filling system includes a bottom-up pattern matching pipeline and a top-down question answering pipeline, with several novel enhancements including statistical answer re-ranking and Markov Logic Networks (MLN) based cross-slot reasoning. From both extraction systems confidence values are produced on various levels: name identification and classification, relation and event labeling and corresponding argument identification and classification.

Based on the assumption that the documents for a given query are topically related, we apply the extraction methods to the each 'super-document' that includes the query and the related documents. As a result we can obtain a knowledge base including entities, relations, events, event chains and coreference links between the query and documents.

This method can be considered as a combination of query expansion and fact retrieval. We not only obtain a 'profile' (potential fact categories) for the query so that

[1] http://www.nist.gov/speech/tests/ace/

2 http://nlp.cs.qc.cuny.edu/kbp/2010/

we can design corresponding templates for abstractive summarization, but also assign weights to sentences including these specific categories of facts.

4 Motivation of Using IE for Summarization

Using the combination of fact types in ACE and KBP, we can cover rich information in news articles. For example, among the 92 TAC queries, 28 queries include explicit ACE events and their corresponding input documents include 2739 event instances. Some queries include specific events such as "*Provide details of the **attacks** on Egypt's Sinai Penninsula resorts targetting Israeli tourists.*", while others only inquire about a general series of events: "*Describe the views and **activities** of John C. Yoo.*"

Previous work has extensively focused on using entity extraction to improve summarization, so we only present some concrete examples of using relations and events to improve summarization quality as follows.

4.1 Relations/Events Can Push Up Relevant Sentences

Traditional sentence ranking methods in summarization used key word matching, and the knowledge acquisition bottleneck still remains due to sparse data. In other words, the training data for similarity matching may not be available for each test instance.

In order to learn a more robust sentence ranker, the method of matching query and sentences should go beyond lexical and syntactic level in order to capture semantic structures. A lot of current extractive summarizers use semantic relations in WordNet ([10]). This approach has two main limitations: (1) It cannot address broader semantic relatedness; (2). It cannot address the semantic relations between two words with different part-of-speech tags. Semantic relation and event classification can provide a more flexible matching framework. Our basic intuition is that a sentence should receive a high rank if it involves many relations and events specified in the query, regardless of the different word forms to indicate such relations and events. For example, for the following query and sentences with high ranks:

[Query]
*London Subway Bombing/Describe the July 7, 2005 **bombings** in **London, England** and the events, casualties and investigation resulting from the attack.*

[***High-Rank Sentence 1***]
*The **attacks**, the deadliest ever carried out on **London** in peacetime, coincided with a summit of the Group of Eight in Gleneagles, Scotland.*

[***High-Rank Sentence 2***]
*A group called Secret al-Qaida Jihad Organization in Europe claimed responsibility, saying the **attacks** were undertaken to avenge **British** involvement in the wars in Afghanistan and Iraq.*

[***High-Rank Sentence 3***]
*The **bomb exploded** in the lead car moments after the train pulled out of the **King's Cross station**, blowing apart the car and making it impossible to reach the dead and injured from the rear.*

In sentences 1 and 2, the baseline summarizer is not able to detect "*attacks*" as the same events as "*bombings*" because they have different lexical forms. The event extraction component, however, predicts "*conflict-attack*" events and labels "*London/British*" as "*place*" arguments in both sentences. This provides us much stronger confidence in increasing the ranks of sentence 1 and 2.

Furthermore, even if the event triggers in sentence 3 "*bomb*" can be matched with "*bombings*" in the query, the baseline summarizer assigns a low weight to sentence 3 because it cannot detect the "located-in" relation between "*King's Cross station*" and "*London*". But the relation extraction component can successfully identify this "*PHYS/Located*" relation from another sentence in the same document set: "*The subway tunnel between **King's Cross** and Russell Square is one of several "deep tubes" bored through **London**'s bedrock and clay more than a century ago*".

4.2 Relations/Events Can Push Down Irrelevant Sentences

On the other hand, relations and events can filter some irrelevant sentences by deep semantic structure analysis. For example,

[Query]
*Judge Joan Lefkow's Family Murdered / Describe the **murders** of **Judge Joan Lefkow**'s **husband and mother**, and the subsequent investigation. Include details about any evidence, witnesses, suspects and motives.*

[***Low-Rank Sentence 4***]
*They remembered that he would sometimes show up at the federal courthouse to take his **wife**, U. S. District Judge **Joan Humphrey Lefkow**, to lunch and brought her flowers.*

The baseline summarizer mistakenly assigns a high rank to sentence 4 because it involves a name "*Joan Humphrey Lefkow*" specified in the query, and "*wife*" can be recognized to match "*husband*" by semantic clusters. However, event extraction can be used to successfully push down this sentence because it does not include any "Conflict-attack (murder)" events.

4.3 Event Coreference Can Remove Redundancy

What we have presented in section 4.1 and 4.2 is advancing summaries in terms of their *Content* quality. Another central track of summarization research is the issue of *readability* – especially how to remove redundancy existing in summaries from multiple documents.

In this paper we propose an approach of using event coreference resolution to reach this goal. Compared to similarity computation methods based on lexical features, our method can detect similar pairs of sentences even if they use completely different expressions. For example, we can fuse the following sentences because they include coreferential "*Conflict-attack*" event instances = Both include indicative words "*blasts/bombings*" and involve "London" as their place arguments:

[***Sentence 5***]
It was the deadliest of the four bomb ***blasts*** *in* ***London*** *last week.*

[***Sentence 6***]
The bus explosion was one of four co-ordinated ***bombings****, the others on* ***London*** *Underground subway trains.*

It is challenging for the baseline summarizer to detect this sentence pair because most words don't overlap.

6 IE-Integrated Summarization

IE provides an effective way of modeling the central information described in the source documents. This model consists of entities, relations and events involving these entities. Even if this model described perfectly such information, it does not tell us what subset of this model should appear in a summary.

The first question we have to tackle is "What is most relevant in IE output?" A baseline estimation method would be to look at the frequency of IE elements in the input and ensure that frequently described events appear in the summary. Another approach would be to build a graph of IE elements, and perform a random walk of this graph to weigh the most relevant nodes. However, both approaches do not account for three factors: relevance prior, coverage and confidence of the extraction.

Another question is "How can we incorporate IE-derived model in a summarization system?" Only considering extractive summarization, approaches vary from scoring sentences directly with supervised or unsupervised relevance assessments, to scoring sub-sentence units and finding best covering sentences. Under those models, IE can be integrated as an extra set of features to characterize either sentences or sub-sentence units. For the purpose of this work, we focus on a simple linear model to blend sentence-level IE scores and a baseline summarizer.

6.1 Approach Overview

Figure 1 depicts the general procedure of our approach to integrate IE results into our baseline summarizer.

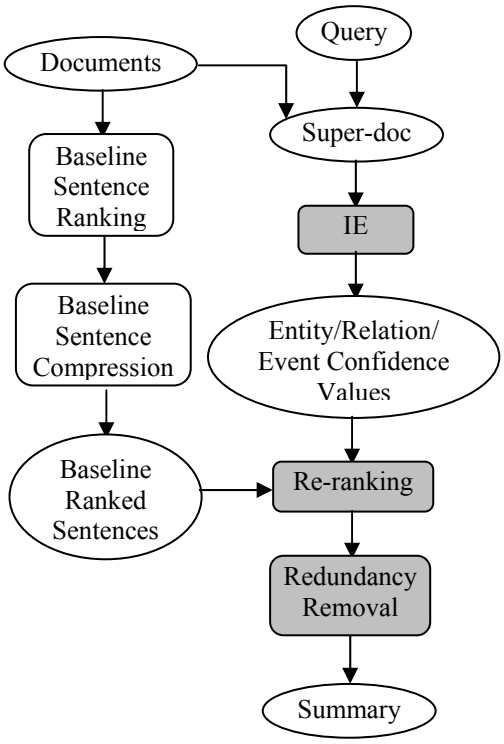

Figure 1. Pipeline of Integrating IE into Summarization

6.2 IE-based Re-Ranking and Redundancy Removal

Because the human summaries are not necessarily created from the original sentences of the input documents, we cannot adopt a supervised learning based re-ranking approach. For each sentence we adjust its rank produced from the baseline summarizer based on IE confidence values.

Each IE component includes a statistical classifier and thus can generate reliable confidence values. For example, for each event mention in *D*, the baseline Maximum Entropy based classifiers produce three types of confidence values:

- *Conf(trigger,etype)*: The probability of a string *trigger* indicating an event mention with type *etype*; if the event mention is produced by pattern matching then assign confidence 1.

- *Conf(arg, etype)*: The probability that a mention *arg* is an argument of some particular event type *etype*.

- *Conf(arg, etype, role)*: If *arg* is an argument with event type *etype*, the probability of *arg* having some particular *role*.

For a given query *Q* and a collection of 10 documents *D* that includes *N* sentences, we generate a summary based on an integrated approach as follows. For any sentence *s* in *D*, we extract various confidence values in Table 1 and combine them to form the final IE confidence for *s*:

$$c_{ie}(s) = \alpha_1 \times \sum_j c_1(s, e_j) + \alpha_2 \times \sum_k c_2(s, r_k) + \alpha_3 \times \sum_l c_3(s, e_l) + \alpha_4 \times \sum_m c_4(s, e_m)$$

Table 1. IE Confidence Values

Confidence	Description
$c_1(s, e_j)$	confidence of *s* including an entity e_j relevant to *Q* (coreferential)
$c_2(s, r_k)$	confidence of *s* including a relation r_k relevant to *Q* (relation type and relation arguments match)
$c_3(s, ev_l)$	confidence of *s* including an event mention ev_k relevant to *Q* (event type and event arguments match)
$c_4(s, evcoref_m)$	confidence of *s* including a link $evcoref_m$ between two coreferential event mentions which are relevant to *Q*

Assuming the ranking confidence from the baseline summarizer is $c_{baseline}(s)$, then we can get the combined weight for *s*:

$$w_{summary}(s) = \lambda_1 \times (c_{baseline}(s) / \sum_{i=1}^{N} c_{baseline}(s_i)) + \lambda_2 \times (c_{ie}(s) / \sum_{i=1}^{N} c_{ie}(s_i))$$

We believe that incorporating these confidence values into a unified re-ranking model can provide a comprehensive representation of the concepts in the source collection of documents. Based on the combined weights, we select top sentences to form a summary according the number of words specified in the task. The parameters α and λ are optimized from a development set. In order to achieve better readability (non-redundancy), we conduct a greedy search through the high-ranked sentences for redundancy removal. If all facts in a sentence pair $<s_i, s_j>$ are determined to be coreferential by our entity and event coreference resolvers, we remove the shorter one.

7 Experimental Results

In this section we present the results of applying IE to improve TAC summarization.

7.1 Data and Evaluation Metrics

We randomly selected 30 topics from TAC 2008 and TAC 2009 summarization task as our development set to optimize parameters and another separate set of 31 topics as our blind test set. The summaries are evaluated automatically with ROUGE-2 and ROUGE-SU4 metrics ([13]). In order to focus more on evaluating the ordering of sentences and coherence across sentences, we extend the length restriction in TAC setting from 100 words to 20 sentences. We also asked 16 human subjects to manually evaluate summaries based on the TAC Responsiveness metric ([7]) consisting of Content and Readability/Fluency measures. In order to compare different methods extensively, we ask the annotators to give a real-value score between [1, 5] (1-Very Poor, 2-Poor, 3-Barely Acceptable, 4-Good, 5-Very Good).

7.2 ROUGE Scores

The parameters α and λ are optimized from a separate development set. We use the following optimized α values: α1=1, α2=2, α3=3, α4=1. Figure 2 presents the effect on ROUGE-2/ROUGE-SU4 scores of varying the IE weight λ, from 0 (baseline summarizer) to 1 (using IE only).

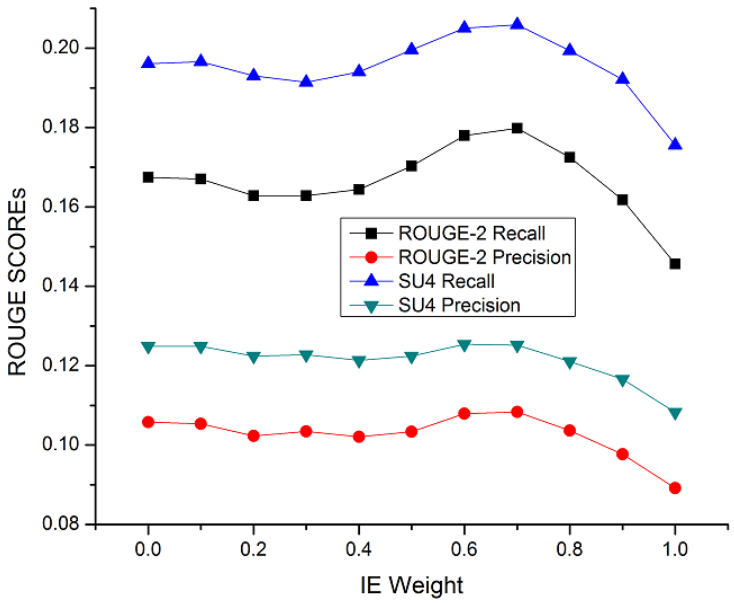

Figure 2. Applying IE to Re-rank the Baseline with Sentence Compression

We can see that our method achieved significant improvement on Recall. When we use λ =0.7, which is also the best weight optimized from the development set, our methods achieved 7.38% relative ROUGE-2 gain. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on

ROUGE scores for these 31 topics. The results show that we can reject the hypothesis that the improvements were random at a 95.7% confidence level. From these curves we can also conclude that using IE results only (λ=1) for sentence ranking produced worse ROUGE scores than the baselines.

7.3 TAC Responsiveness Scores

Table 2 presents the average scores across all topics based on manual evaluation using TAC Responsiveness metrics.

Table 2. TAC Responsiveness Comparison

Method	Content	Readability	Responsiveness
Baseline	3.11	3.56	3.39
IE-Integrated	3.89	3.67	3.61

Table 2 shows that our IE-integrated method received much better Content scores based on human assessment. For example, for the query "Provide details of the kidnapping of journalist Jill Carroll in Baghdad and the efforts to secure her release", the baseline summarizer received a score '2' because of mis-match between 'kidnapping' in the query and the 'arrest' events involving other person and place arguments in the source documents. In contrast, our method received a score '4', because of the effective integration of 'kidnap' event detection results to re-rank sentences. Furthermore, according to the user feedback, our method produced fewer redundant sentences for most topics.

7.4 Discussion

Error analysis shows that for 3 topics IE had negative impact because of incorrect event categorization for the queries, and missing/spurious extraction errors. For example, for the query "*BTK/Track the efforts to identify the serial killer BTK and bring him to justice.*", IE mistakenly recognized 'Justice' as the main event type while missed a more important event type 'Investigation' which was not defined in the 33 event types. In these and other cases, we could apply salience detection to assign weights to different facts types in the query. Nevertheless, as the above results indicate, the rewards of using the IE information outweigh the risks.

8 Related Work

Our work is a re-visit on the idea of exploiting IE results to improve multi-document summarization proposed by Radev et al. ([2]) and White et al. ([3]). In ([2]), IE results such as entities and MUC events are combined with natural language generation techniques in summarization. White et al. ([3]) improved Radev et al.'s method by

summarizing larger input documents based on relevant content selection and sentence extraction. They also formally evaluated the performance of this idea. More recently, Filatova and Hatzivassiloglou ([14]) considered the contexts involving any pair of names as general 'events' and used them to improve extractive summarization. Vanderwende et al. ([15]) explored an event-centric approach and generated summaries based on extracting and merging portions of logical forms. Biadsy et al. ([16]) exploited entity and time facts extracted from IE to improve sentence extraction for biographical summaries. Hachey ([11]) used generic relations to improve extractive summarization and remove redundancy. Compared to these previous methods, we extend the usage of IE from single template to much more complete relation/event types. To the best of our knowledge our approach is the first work to use the information extracted from KBP project in summarization and apply event coreference resolution to remove summary redundancy.

In addition, our work is related to the summarization research that incorporates semantic role labeling (SRL) results (e.g. [12, 17]). SRL has a higher coverage on event categories than IE, while IE can provide additional annotations such as entity resolution and event resolution which are beneficial to summarization.

Our approach of selecting informative facts is also similar to defining Summarization Content Units (SCUs) in the Pyramid Approach ([18]) because both methods aim to maximize the coverage of logical 'concepts' in summaries..

9 Conclusion

We investigated the once-popular IE-driven summarization approaches in a wider IE paradigm. We demonstrated that a simple re-ranking approach can achieve improvement over a high-performing extractive summarizer. We expect that as IE is further developed to achieve higher performance in wider domains, the summarization task can benefit more from extended semantic frames.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Ralph Grishman, Jerry Hobbs, Eduard Hovy, Antonio Sanfilippo and Yorick Wilks. 1997. Cross-lingual Information Extraction and Automated Text Summarization. *Linguistica Computazionale, Volume XIV-XV.*
2. Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3):469–500.
3. Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce and Kiri Wagstaff. 2001. Multidocument Summarization via Information Extraction. *Proc. HLT 2001.*
4. Ralph Grishman and Beth Sundheim: Message Understanding Conference - 6: A Brief History. *Proc. COLING 1996.*.
5. Heng Ji, Ralph Grishman, Zheng Chen and Prashant Gupta. 2009. Cross-document Event Extraction, Ranking and Tracking. *Proc. RANLP 2009*.
6. Heng Ji, Zheng Chen, Jonathan Feldman, Antonio Gonzalez, Ralph Grishman and Vivek Upadhyay. 2010. Utility Evaluation of Cross-document Information Extraction. *Proc. NAACL/HLT 2010.*
7. Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 Summarization Track. *Proc. TAC 2009*.
8. Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, Shasha Xie. 2009. The ICSI/UTD Summarization System at TAC 2009. *Proc. TAC 2009*.
9. Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Marissa Passantino and Heng Ji. 2011. Top-down and Bottom-up: A Combined Approach to Slot Filling. *Proc. AIRS 2011.*
10. Christiane Fellbaum (Ed.). 1998. WordNet*: An Electronic Lexical Database. Cambridge*, MA: The MIT Press.
11. Ben Hachey. 2009. Multi-Document Summarisation Using Generic Relation Extraction. *Proc. EMNLP 2009*.
12. Gabor Melli, Yang Wang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar and Fred Popowich. 2005. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task. *Proc. DUC workshop 2005.*
13. Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. *Proc. HLT-NAACL 2003*.
14. E. Filatova and V. Hatzivassiloglou, 2004. A Formal Model for Information Selection in Multi-Sentence Text Extraction. *Proc. COLING 2004*.
15. Lucy Vanderwende, Michele Banko and Arul Menezes. 2004. Event-Centric Summary Generation. *Proc. DUC 2004.*
16. Fadi Biadsy, Julia Hirschberg and Elena Filatova. 2008. An Unsupervised Approach to Biography Production using Wikipedia. *Proc. ACL 2008*.
17. Gabor Melli, Zhongmin Shi, Yang Wang, Yudong Liu, Anoop Sarkar and Fred Popowich. 2006. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2006 Summarization Task. *Proc. DUC 2006*.
18. Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method*. Proc. NAACL 2004*.
19. Elena Filatova and Vasileios Hatsivassiloglou. 2004. Event-based extractive summarization. *Proc. ACL Workshop on Summarization*.