# Improving Update Summarization via Supervised ILP and Sentence Reranking

**Chen Li**[1], **Yang Liu**[1], **Lin Zhao**[2]
[1] Computer Science Department, The University of Texas at Dallas
Richardson, Texas 75080, USA
[2] Research and Technology Center, Robert Bosch LLC
Palo Alto, California 94304, USA
{chenli,yangl@hlt.utdallas.edu}
{lin.zhao@us.bosch.com}

## Abstract

Integer Linear Programming (ILP) based summarization methods have been widely adopted recently because of their state-of-the-art performance. This paper proposes two new modifications in this framework for update summarization. Our key idea is to use discriminative models with a set of features to measure both the salience and the novelty of words and sentences. First, these features are used in a supervised model to predict the weights of the concepts used in the ILP model. Second, we generate preliminary sentence candidates in the ILP model and then rerank them using sentence level features. We evaluate our method on different TAC update summarization data sets, and the results show that our system performs competitively compared to the best TAC systems based on the ROUGE evaluation metric.

## 1 Introduction

Update summarization has attracted significant research focus recently. Different from generic extractive summarization, update summarization assumes that users already have some information about a given topic from an old data set, and thus for a new data set the system aims to generate a summary that contains as much novel information as possible. This task was first introduced at DUC 2007 and then continued until TAC 2011. It is very useful to chronological events in real applications.

Most basic update summarization methods are variants of multi-document summarization methods, with some consideration of the difference between the earlier and later document sets (Boudin et al., 2008; Fisher and Roark, 2008; Long et al., 2010; Bysani, 2010). One important line is to use graph-based co-ranking. They rank the sentences in the earlier and later document sets simultaneously by considering the sentence relationship. For example, Li et al. (2008) was inspired by the intuition that "a sentence receives a positive influence from the sentences that correlate to it in the same collection, whereas receives a negative influence from the sentences that correlates to it in the different (or previously read) collection', and proposed a graph based sentence ranking algorithm for update summarization. Wan (2012) integrated two co-ranking processes by adding some strict constraints, which led to more accurate computation of sentences' scores for update summarization. A similar method was also applied earlier by (Wan et al., 2011) for multilingual news summarization. In addition, generative models, such as topic models, have also been adopted for this task. For example, Delort and Alfonseca (2012) proposed a novel nonparametric Bayesian approach, a variant of Latent Dirichlet Allocation (LDA), aiming to distinguish between common information and novel information. Li et al. (2012) borrowed the idea of evolutionary clustering and proposed a three-level HDP (Hierarchical Dirichlet Process) model to represent the diversity and commonality between aspects discovered from two different document data sets.

One of the most competitive summarization methods is based on Integer Linear Programming (ILP). It has been widely adopted in the generic summarization task (Martins and Smith, 2009; Berg-Kirkpatrick et al., 2011; Woodsend and Lapata, 2012; Li et al., 2013a; Li et al., 2013b; Li et al., 2014). In this paper, we use the ILP summarization

*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1317–1322,
Denver, Colorado, May 31 – June 5, 2015.

framework for the update summarization task, and make improvement from two aspects, with the goal to more discriminatively represent both the salience and novelty of words and sentences. First, we use supervised models and a rich set of features to learn the weights for the bigram concepts used in the ILP model. Second, we design a sentence reranking component to score the summary candidate sentences generated by the ILP model. This second reranking approach allows us to explicitly model a sentence's importance and novelty, which complements the bigram centric view in the first step of ILP sentence selection. Our experimental results on multiple TAC data sets demonstrate the effectiveness of our proposed method.

## 2 Proposed Update Summarization System

### 2.1 ILP Framework for Summarization

The core idea of the ILP based summarization method is to select the summary sentences by maximizing the sum of the weights of the language concepts that appear in the summary. Bigrams are often used as the language concepts in this method. Gillick et al. (2009) stated that the bigrams gave consistently better performance than unigrams or trigrams for a variety of ROUGE measures. The association between the language concepts and sentences serves as the constraints. This ILP method is formally represented as below (see (Gillick et al., 2009) for more details):

$$\begin{array}{lll} \max & \sum_i w_i c_i & \\ s.t. & c_i \in \{0,1\} \; \forall i & s_j \in \{0,1\} \; \forall j \\ & s_j Occ_{ij} \le c_i & \sum\limits_j s_j Occ_{ij} \ge c_i \\ & \sum_j l_j s_j \le L & \end{array} \tag{1}$$

where $c_i$ and $s_j$ are binary variables that indicate the presence of a concept and a sentence respectively. $l_j$ is the sentence length and $L$ is the maximum length (word number) of the generated summary. $w_i$ is a concept's weight and $Occ_{ij}$ means the occurrence of concept $i$ in sentence $j$. The first two inequalities associate the sentences and concepts. They ensure that selecting a sentence leads to the selection of all the concepts it contains, and selecting a concept only happens when it is present in at least one of the selected sentences.

### 2.2 Bigrams Weighting for Salience and Novelty

In the above ILP-based summarization method, how to determine the concepts and measure their weights is the key factor impacting the system performance. Intuitively, if we can successfully identify the important key bigrams used in the ILP system, or assign large weights to those important bigrams, the generated summary sentences will contain as many important bigrams as possible, and thus resulting in better summarization performance. The oracle experiment in (Gillick et al., 2008) showed that if they use the bigrams extracted from the human written summaries as the input of the ILP system, much better ROUGE scores can be obtained than using the automatically selected bigrams, suggesting the importance of using the right concepts. (Gillick et al., 2009) used document frequency as the weight of a bigram. They also provided some justification for document frequency as a weighting function in that paper.

For update summarization, intuitively we need to not only identify the salience of the bigram, but also incorporate bigrams' novelty in their weights. Therefore, only using the document frequency as the weight in the objective function is insufficient. We thus propose to use a supervised framework for the bigram weight estimation in the ILP model. The new objective function is:

$$\max \quad \sum_i (\theta \cdot \mathbf{f}(b_i))\, c_i \tag{2}$$

We replace the heuristic $w_i$ in Formula (1) with a feature based one: $f(b_i)$ represents the features for a bigram $b_i$, and $\theta$ is a weight vector for these features. Constraints remain the same as before in the ILP method.

There are two kinds of features for each bigram: one set is related to the bigrams themselves; the other set is related to the sentences containing the bigram. Table 1 shows the features we design. For both the bigram and the sentence level features, we separate the features based on whether they represent the importance or the novelty. For

Feature 8 and 17, the summary is generated by a general unsupervised ILP-based summarization system from the given old data set. The idea of Feature 9 and 10 was first introduced by (Bysani, 2010); here we applied it to bigrams. $df_{max}$ is the number of documents in the data set (10 in the TAC data), which can be thought of the maximum value of document frequency for a bigram. Feature 11 is interpolated n-gram document frequency, which was first introduced by (Ng et al., 2012): $\frac{\alpha \sum_{w_u \in S} df_{new}(w_u) + (1-\alpha) \sum_{w_b \in S} df_{new}(w_b)}{|S|}$, where $w_u$ and $w_b$ are unigrams and bigrams respectively in sentence $S$. Feature 18 and 19 are variants of Features 11, where instead of document frequency ($df$ in the formula above), bigram and unigram's novelty and uniqueness values are used. Among these features, the feature values of feature 4, 5 and 6 are discrete. In this study, we discretized all the other continuous values into ten categories according to the value range in the training data.

To train the model (feature weights), we use the average perceptron strategy (Collins, 2002) to update the feature weights whenever the hypothesis by the ILP decoding process is incorrect. Binary class labels are used for bigrams in the learning process, that is, we only consider whether a bigram is in the system generated summary or human summaries, not their term or document frequency. We use a fixed learning rate (0.1) in training.

### 2.3 Sentence Reranking on ILP Results

In the ILP method, sentence selection is done by considering the concepts that a sentence contains. It is difficult to add indicative features in this framework to explicitly represent the sentence's salience, and more importantly, its novelty for the update summarization task. This information is only captured by the weights of the bigrams using the method described above. Therefore, we propose to use a two-step approach, where an initial ILP module first selects some sentences and then a reranking module uses sentence level features to rerank them to generate the final summary. We expect this step of modeling sentences directly can complement the bigram centric view in the first ILP summarization module.

**Bigram Level Features**

**Importance Related Features**

1. $df_{new}(b)$: document frequency in new data set

2. normalized term frequency in all filtered relevant sentences[1]

3. sentence frequency in all relevant sentences

4. do bigram words appear in topic's query ?

5. is the bigram in the first 1/2/3 position of that sentence?

6. is the bigram in the last 1/2/3 position of that sentence?

**Novelty Related Features**

7. $df_{old}(b)$: document frequency in old data set

8. normalized term frequency in the summary from old data set

9. bigram novelty value $n(b) = \frac{df_{new}(b)}{df_{old}(b) + df_{max}}$

10. bigram uniqueness value $u(b) = 0$ if $df_{old}(b) > 0$; otherwise $u(b) = \frac{df_{new}(b)}{df_{max}}$

**Sentence Level Features**

**Importance Related Features**

11. interpolated n-gram document frequency

12. sentence position in that document

13. is the sentence in the first 1/2/3 position in that document?

14. is the sentence in the last 1/2/3 position in that document?

15. sentence length

16. sentence similarity with topic's query

**Novelty Related Features**

17. sentence similarity with the summary from old data set

18. interpolated n-gram novelty

19. interpolated n-gram uniqueness

Table 1: Features in the supervised ILP model for weighting bigrams.

For the first step, we use the ILP framework with our supervised bigram weighting method to obtain a summary of $N$ words ($N$ is greater than the required summary length $L$). Note that the ILP model selects these output sentences as a set that optimizes the objective function, and there are no scores for each individual sentence. Second, we use sentence level features listed in Table 1 to rerank the can-

[1] Note that we do not use all the sentences in the ILP module. The 'relevant' sentences are those that have at least one bigram with document frequency larger than or equal to three.

didate sentences. This is expected to better evaluate the salience and the novelty of the sentences. We use a regression model (SVR) for this reranking purpose. When training the model, a sentence's ROUGE2 score compared with the human generated summary is used as the regression target. After reranking, we just select the top sentences that satisfy the length constraint to form the final summary. In this work we do not use any redundancy removal (e.g., MMR method). This is because the ILP decoding process tries to find a global optimal set maximizing the concept coverage, subject to the length constraint, and thus already considers redundancy among sentences. Typically when the initial set (i.e., the output from the first ILP step) is not too big, redundancy is not a big problem.

## 3 Experiments and Results

### 3.1 Data and Experiment Setup

We evaluate our methods using several recent TAC data sets, from 2008 to 2011. Every topic has two sets of 10 documents (Set A and B). The update task aims to create a 100-word summary from Set B given a topic query and Set A. When evaluating on one year's data, we use the data from the other three years as the training set. This applies to both the supervised ILP method and the sentence reranking regression model. All the summaries are evaluated using ROUGE (Lin, 2004). An academic free solver[2] does all the ILP decoding and libsvm[3] is used for SVR implementation.

### 3.2 Results

Table 2 and Table 3 show the R2 and R-SU4 values on different TAC data sets for the following systems.

- ILP baseline. This is the unsupervised ILP-based summarization system (Gillick et al., 2009), in which only bigrams with document frequency greater than 2 are used in the ILP summarization process, and weight $w_i$ is the document frequency of that bigram.
- TAC best. This is the best result in the TAC update summarization evaluation.[4] Note that there is limited research on update summarization and we cannot find better published results for these data sets than the TAC best systems.
- Supervised ILP. This is our supervised ILP method where bigram weights are learned discriminately. It is the one-step system that generates the summary with the target length. We use the same bigram set as the ILP baseline system. For this method, we show results using different features: only using the importance features; and using all the features. This is used to evaluate the impact of the novelty features on the update summarization task.
- Two-step method: supervised ILP followed by sentence reranking. We generate 200 (value of $N$) words summary in the ILP system. Two different configurations are also used: with and without the sentence novelty features in the sentence ranking module. All the features (including the novelty features) are used in the ILP pre-selection step.
- Sentence ranking without ILP. In this experiment, we do not use the ILP summarization module to generate candidate sentences first, but just apply sentence ranking to the entire data set. Then MMR is leveraged to select the final summary sentences. Again, we present results using different feature sets.

We can see from the tables that the supervised ILP model outperforms the unsupervised one. After including the novelty related features, the model can assign higher weights for the bigrams with novel information, resulting in improved summarization performance. There is further improvement when using our 2-step approach with the sentence reranking model. Our proposed method (ILP followed by sentence reranking, and using all the features) outperforms the TAC best result in 2010 and 2011, and also yields competitive results in the other data sets. The gain of ROUGE-2 of our proposed system compared with the ILP baseline is statistically significant based on ROUGE's 95% confidence. When using sentence ranking on the entire document set, without the ILP pre-selection step, its performance is worse than our proposed method. This shows the benefit of doing pre-selection using the ILP module. Finally,

[2] http://www.gurobi.com

[3] http://www.csie.ntu.edu.tw/c̃jlin/libsvm/

[4] The ID of the TAC best system from 2008 to 2011 is 14,40,16 and 43.

|  | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|
| ILP Baseline | 8.55 | 8.84 | 7.04 | 8.63 |
| TAC Best | 10.10 | 10.41 | 8.00 | 9.58 |
| Supervised ILP |  |  |  |  |
| w/o novelty features | 9.18 | 9.06 | 7.39 | 9.20 |
| w all features | 9.4 | 9.28 | 7.76 | 9.46 |
| 2-step: Supervised ILP + Sentence Ranking |  |  |  |  |
| w/o novelty features | 9.65 | 9.47 | 7.97 | 9.70 |
| w all features | 9.99 | 9.61 | 8.11 | 9.99 |
| Sentence Ranking w/o ILP |  |  |  |  |
| w/o novelty features | 9.25 | 9.10 | 7.41 | 9.18 |
| w all features | 9.42 | 9.32 | 7.70 | 9.43 |

Table 2: ROUGE-2 results on TAC 2008-2011 data.

|  | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|
| ILP Baseline | 12.17 | 12.54 | 10.57 | 12.01 |
| NIST Best | 13.66 | 13.95 | 11.97 | 13.08 |
| Supervised ILP |  |  |  |  |
| w/o novelty features | 12.57 | 12.94 | 11.01 | 12.76 |
| w all features | 12.78 | 13.21 | 11.61 | 12.95 |
| 2-step: Supervised ILP + Sentence Ranking |  |  |  |  |
| w/o novelty features | 13.10 | 13.65 | 11.98 | 13.24 |
| w all features | 13.61 | 13.77 | 12.20 | 13.42 |
| Sentence Ranking w/o ILP |  |  |  |  |
| w/o novelty features | 12.60 | 12.99 | 11.25 | 12.73 |
| w all features | 12.85 | 13.31 | 11.50 | 12.90 |

Table 3: ROUGE-SU4 results on TAC 2008-2011 data.

for all the methods, adding the novelty related features always performs better than that without them, proving the effect of our novelty features for update summarization.

Lastly we evaluate the effect of the summary length from the ILP module on the two-step summarization systems. Figure 1 shows the performance when $N$ changes from 150 to 400. We can see that there is some difference in the patterns for different data sets, and the best results are obtained when $N$ is around 150 to 250. When the first ILP module produces many sentence candidates, it is likely that there is redundancy among them. In this case, redundancy removal approaches such as MMR need to be used to generate the final summary. In addition, for a large candidate set, our current regression model also faces some challenges due to its limited features used in sentence reranking. Addressing these problems is our future work.

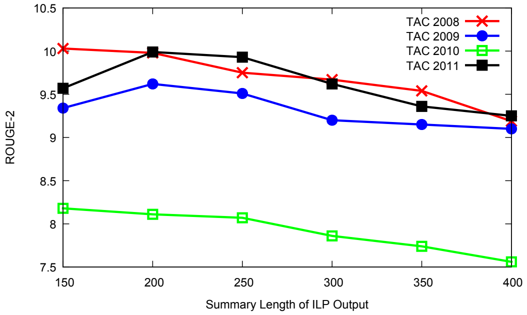

Figure 1: ROUGE-2 results when varying the output length for the first ILP selection step.

## 4 Conclusions

In this paper, we adopt the supervised ILP framework for the update summarization task. A set of rich features are used to measure the importance and novelty of the bigram concepts used in the ILP model. In addition, we proposed a re-selection component to rank candidate sentences generated by the ILP model based on sentence level features. Our experiment results show that our features and the reranking procedure both help improve the summarization performance. This pilot research points out new directions for generic or update summarization based on the ILP framework.

## Acknowledgments

We thank the anonymous reviewers for their detailed and insightful comments on earlier drafts of this paper. The work is partially supported by NSF award IIS-0845484 and DARPA Contract No. FA8750-13-2-0041. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views of the funding agencies.

## References

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of ACL*.

Florian Boudin, Marc El-Bèze, and Juan-Manuel Torres-Moreno. 2008. The LIA update summarization systems at tac-2008. In *Proceedings of TAC*.

Praveen Bysani. 2010. Detecting novelty in the context of progressive summarization. In *Proceedings of the NAACL Student Research Workshop*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Jean-Yves Delort and Enrique Alfonseca. 2012. Dualsum: a topic-model based approach for update summarization. In *Proceedings of EACL*.

Seeger Fisher and Brian Roark. 2008. Query-focused supervised sentence ranking for update summaries. In *Proceeding of TAC*.

Dan Gillick, Benoit Favre, and Dilek Hakkani-tur. 2008. The ICSI summarization system at tac 2008. In *Proceedings of TAC*.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The ICSI/UTD summarization system at tac 2009. In *Proceedings of TAC*.

Wenjie Li, Furu Wei, Qin Lu, and Yanxiang He. 2008. PNR2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of Coling*.

Jiwei Li, Sujian Li, Xun Wang, Ye Tian, and Baobao Chang. 2012. Update summarization using a multi-level hierarchical dirichlet process model. In *Proceedings of COLING*.

Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. 2013a. Document summarization via guided sentence compression. In *Proceedings of the EMNLP*.

Chen Li, Xian Qian, and Yang Liu. 2013b. Using supervised bigram-based ilp for extractive summarization. In *Proceedings of ACL*.

Chen Li, Yang Liu, Fei Liu, Lin Zhao, and Fuliang Weng. 2014. Improving multi-documents summarization by sentence compression based on expanded constituent parse trees. In *Proceedings of EMNLP*.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of ACL*.

Chong Long, Minlie Huang, and Xiaoyan Zhu. 2010. Summarizing multidocuments by information distance. In *Proceedings of TAC*.

Andre F. T. Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the ACL Workshop on Integer Linear Programming for Natural Language Processing*.

Jun-Ping Ng, Praveen Bysani, Ziheng Lin, Min-Yen Kan, and Chew-Lim Tan. 2012. Exploiting category-specific information for multi-document summarization. In *Proceedings of COLING*.

Xiaojun Wan, Houping Jia, Shanshan Huang, and Jianguo Xiao. 2011. Summarizing the differences in multilingual news. In *Proceedings of SIGIR*.

Xiaojun Wan. 2012. Update summarization based on co-ranking with constraints. In *Proceedings of COLING*.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of EMNLP-CoNLL*.