Towards Automatic Generation of Gene Summary

Feng Jin
Dept. Computer Science and Technology
Tsinghua University
Beijing 100084, China
jinfengfeng@gmail.com

Minlie Huang
Dept. Computer Science and Technology
Tsinghua University
Beijing 100084, China
aihuang@tsinghua.edu.cn

Zhiyong Lu
National Center for Biotechnology Information
National Library of Medicine
Bethesda, 20894, USA
luzh@ncbi.nlm.nih.gov

Xiaoyan Zhu
Dept. Computer Science and Technology
Tsinghua University
Beijing 100084, China
zxy-dcs@tsinghua.edu.cn

Abstract

In this paper we present an extractive system that automatically generates gene summaries from the biomedical literature. The proposed text summarization system selects and ranks sentences from multiple MEDLINE abstracts by exploiting gene-specific information and similarity relationships between sentences. We evaluate our system on a large dataset of 7,294 human genes and 187,628 MEDLINE abstracts using Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a widely used automatic evaluation metric in the text summarization community. Two baseline methods are used for comparison. Experimental results show that our system significantly outperforms the other two methods with regard to all ROUGE metrics. A demo website of our system is freely accessible at http://60.195.250.72/onbires/summary.jsp.

1 Introduction

Entrez Gene is a database for gene-centric information maintained at the National Center for Biotechnology Information (NCBI). It includes genes from completely sequenced genomes (e.g. Homo sapiens). An important part of a gene record is the summary field (shown in Table 1), which is a small piece of text that provides a quick synopsis of what is known about the gene, the function of its encoded protein or RNA products, disease associations, genetic interactions, etc. The summary field, when available, can help biologists to understand the target gene quickly by compressing a huge amount of knowledge from many papers to a small piece of text. At present, gene summaries are generated manually by the National Library of Medicine (NLM) curators, a time- and labor-intensive process. A previous study has concluded that manual curation is not sufficient for annotation of genomic databases (Baumgartner et al., 2007). Indeed, of the 5 million genes currently in Entrez Gene, only about 20,000 genes have a corresponding summary. Even in humans, arguably the most important species, the coverage is modest: only 26% of human genes are curated in this regard. The goal of this work is to develop and evaluate computational techniques towards automatic generation of gene summaries.

To this end, we developed a text summarization system that takes as input MEDLINE documents related to a given target gene and outputs a small set of genic information rich sentences. Specifically, it first preprocesses and filters sentences that do

Proceedings of the Workshop on BioNLP, pages 97–105,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

Gene	*Number of Abstracts*	*GO terms*	*Human-writtenSummary*
EFEMP1	26	calcium ion binding protein binding extracellular region proteinaceous extracellular matrix	This gene spans approximately 18 kb of genomic DNA and consists of 12 exons. Alternative splice patterns in the 5\' UTR result in three transcript variants encoding the same extracellular matrix protein. Mutations in this gene are associated with Doyne honeycomb retinal dystrophy.
IL20RA	15	blood coagulation receptor activity integral to membrane membrane	The protein encoded by this gene is a receptor for interleukin 20 (IL20), a cytokine that may be involved in epidermal function. The receptor of IL20 is a heterodimeric receptor complex consisting of this protein and interleukin 20 receptor beta (IL20B). This gene and IL20B are highly expressed in skin. The expression of both genes is found to be upregulated in Psoriasis.

Table1. Two examples of human-written gene summaries

not include enough informative words for gene summaries. Next, the remaining sentences are ranked by the sum of two individual scores: a) an authority score from a lexical PageRank algorithm (Erkan and Radev, 2004) and b) a similarity score between the sentence and the Gene Ontology (GO) terms with which the gene is annotated (To date, over 190,000 genes have two or more associated GO terms). Finally, redundant sentences are removed and top ranked sentences are nominated for the target gene.

In order to evaluate our system, we assembled a gold standard dataset consisting of handwritten summaries for 7,294 human genes and conducted an *intrinsic* evaluation by measuring the amount of overlap between the machine-selected sentences and human-written summaries. Our metric for the evaluation was ROUGE[1], a widely used intrinsic summarization evaluation metric.

2 Related Work

Summarization systems aim to extract salient text fragments, especially sentences, from the original documents to form a summary. A number of methods for sentence scoring and ranking have been developed. Approaches based on sentence position (Edmundson, 1969), cue phrase (McKeown and Radev, 1995), word frequency (Teufel and Moens, 1997), and discourse segmentation (Boguraev and Kennedy, 1997) have been reported. Radev et al. (Radev et al., 2004) developed an extractive multi-document summarizer, MEAD, which extracts a summary from multiple documents based on the document cluster centroid, position and first-sentence overlap. Recently, graph-based ranking methods, such as LexPageRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004), have been proposed for multi-document summarization. Similar to the original PageRank algorithm, these methods make use of similarity relationships between sentences and then rank sentences according to the "votes" or "recommendations" from their neighboring sentences.

Lin and Hovy (2000) first introduced topic signatures which are topic relevant terms for summarization. Afterwards, this technique was successfully used in a number of summarization systems (Hickl et al., 2007, Gupta and Nenkova et al., 2007). In order to improve sentence selection, we adopted the idea in a similar way to identify terms that tend to appear frequently in gene summaries and subsequently filter sentences that include none or few such terms.

Compared with newswire document summarization, much less attention has been paid to summarizing MEDLINE documents for genic information. Ling et al. (Ling et al., 2006 and 2007) presented an automatic gene summary generation system that constructs a summary based on six aspects of a gene, such as gene products, mutant phenotype, etc. In their system, sentences were ranked according to a) the relevance to each category (namely the aspect), b) the relevance to the document where they are from; and c) the position where sentences are located. Although the system performed well on a small group of genes (10~20 genes) from Flybase, their method relied heavily on high-quality training data that is often hard to obtain in practice.

Yang et al. reported a system (Yang et al., 2007 and 2009) that produces gene summaries by focusing on gene sets from microarray experiments. Their system first clustered gene set into functional related groups based on free text, Medical Subject Headings (MeSH®) and Gene Ontology (GO) features. Then, an extractive summary was generated for each gene following the Edmundson paradigm

[1] http://haydn.isi.edu/ROUGE/

(Edmundson, 1969). Yang et al. also presented evaluation results based on human ratings of eight gene summaries.

Another related work is the second task of Text REtrieval Conference[2] (TREC) 2003 Genomics Track. Participants in the track were required to extract GeneRIFs from MEDLINE abstracts (Hersh and Bhupatiraju, 2003). Many teams approached the task as a sentence classification problem using GeneRIFs in the Entrez database as training data (Bhalotia et al., 2003; Jelier et al., 2003). This task has also been approached as a single document summarization problem (Lu et al., 2006).

The gene summarization work presented here differs from the TREC task in that it deals with multiple documents. In contrast to the previously described systems for gene summarization, our approach has three novel features. First, we are able to summarize all aspects of gene-specific information as opposed to a limited number of predetermined aspects. Second, we exploit a lexical PageRank algorithm to establish similarity relationships between sentences. The importance of a sentence is based not only on the sentence itself, but also on its neighbors in a graph representation. Finally, we conducted an intrinsic evaluation on a large publicly available dataset. The gold standard assembled in this work makes it possible for comparisons between different gene summarization systems without human judgments.

3 Method

To determine if a sentence is extract worthy, we consider three different aspects: (1) the number of salient or informative words that are frequently used by human curators for writing gene summaries; (2) the relative importance of a sentence to be included in a gene summary; (3) the gene-specific information that is unique between different genes.

Specifically, we look for signature terms in handwritten summaries for the first aspect. Ideally, computer generated summaries should resemble handwritten summaries. Thus the terms used by human curators should also occur frequently in automatically generated summaries. In this regard, we use a method similar to Lin and Hovy (2000) to identify signature terms and subsequently use them to discard sentences that contain none or few such terms. For the second aspect, we adopt a lexical PageRank method to compute the sentence importance with a graph representation. For the last aspect, we treat each gene as having its own properties that distinguish it from others. To reflect such individual differences in the machine-generated summaries, we exploit a gene's GO annotations as a surrogate for its unique properties and look for their occurrence in abstract sentences.

Our gene summarization system consists of three components: a preprocessing module, a sentence ranking module, and a redundancy removal and summary generation module. Given a target gene, the preprocessing module retrieves corresponding MEDLINE abstracts and GO terms according to the gene2pubmed and gene2go data provided by Entrez Gene. Then the abstracts are split into sentences by the MEDLINE sentence splitter in the LingPipe[3] toolkit. The sentence ranking module takes these as input and first filters out some non-informative sentences. The remaining sentences are then scored according to a linear combination of the PageRank score and GO relevance score. Finally, a gene summary is generated after redundant sentences are removed. The system is illustrated in Figure 1 and is described in more detail in the following sections.

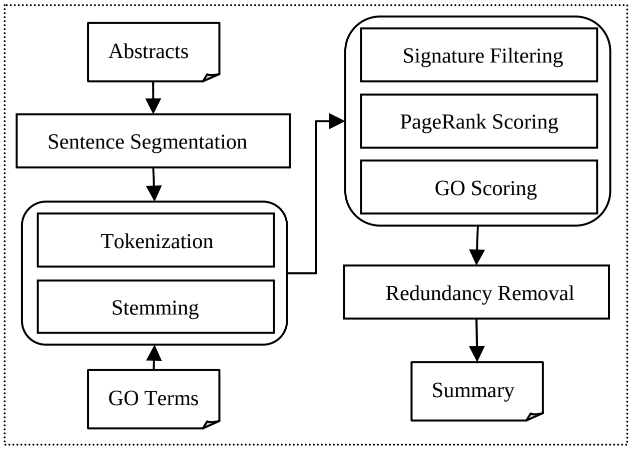

Figure 1. System overview

3.1 Signature Terms Extraction

There are signature terms for different topic texts (Lin and Hovy, 2000). For example, terms such as *eat*, *menu* and *fork* that occur frequently in a corpus may signify that the corpus is likely to be

[2] http://ir.ohsu.edu/genomics/

[3] http://alias-i.com/lingpipe/

about cooking or restaurants. Similarly, there are signature terms for gene summaries.

We use the Pearson’s chi-square test (Manning and Schütze, 1999) to extract topic signature terms from a set of handwritten summaries by comparing the occurrence of terms in the handwritten summaries with that of randomly selected MEDLINE abstracts. Let R denote the set of handwritten summaries and $\tilde{R}$ denote the set of randomly selected abstracts from MEDLINE. The null hypothesis and alternative hypothesis are as follows:

$$\mathrm{H}_0: \quad P(t_i \mid R) = p = P(t_i \mid \tilde{R})$$

$$\mathrm{H}_1: \quad P(t_i \mid R) = p_1 \neq p_2 = P(t_i \mid \tilde{R})$$

The null hypothesis says that the term t_i appears in R and in $\tilde{R}$ with an equal probability and t_i is independent from R. In contrast, the alternative hypothesis says that the term t_i is correlated with R. We construct the following 2-by-2 contingency table:

	R	$\tilde{R}$
t_i	O_{11}	O_{12}
$\tilde{t}_i$	O_{21}	O_{22}

Table 2. Contingency table for the chi-square test.

where

O_{11} : the frequency of term t_i occurring in R;

O_{12} : the frequency of t_i occurring in $\tilde{R}$;

O_{21} : the frequency of term $\tilde{t}_i \neq t_i$ occurring in R;

O_{22} : the frequency of $\tilde{t}_i$ in $\tilde{R}$.

Then the Pearson’s chi-square statistic is computed by

$$X^2 = \sum_{i,j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency.

In our experiments, the significance level is set to 0.001, thus the corresponding chi-square value is 10.83. Terms with X^2 value above 10.83 would be selected as signature terms. In total, we obtained 1,169 unigram terms. The top ranked (by X^2 value) signature terms are listed in Table 3. Given the set of signature terms, sentences containing less than 3 signature terms are discarded. This parameter was determined empirically during the system development.

protein	member	receptor
gene	variant	isoform
encode	domain	alternative
family	splice	bind
transcription	subunit	involve

Table 3. A sample of unigram topic signature terms.

3.2 Lexical PageRank Scoring

The lexical PageRank algorithm makes use of the similarity between sentences and ranks them by how similar a sentence is to all other sentences. It originates from the original PageRank algorithm (Page et al., 1998) that is based on the following two hypotheses:

(1) A web page is important if it is linked by many other pages.

(2) A web page is important if it is linked by important pages.

The algorithm views the entire internet as a large graph in which a web page is a vertex and a directed edge is connected according to the linkage. The salience of a vertex can be computed by a random walk on the graph. Such graph-based methods have been widely adapted to such Natural Language Processing (NLP) problems as text summarization and word sense disambiguation. The advantage of such graph-based methods is obvious: the importance of a vertex is not only decided by itself, but also by its neighbors in a graph representation. The random walk on a graph can imply more global dependence than other methods. Our PageRank scoring method consists of two steps: constructing the sentence graph and computing the salience score for each vertex of the graph.

Let $S = \{s_i \mid 1 \leq i \leq N\}$ be the sentence collection containing all the sentences to be summarized. According to the vector space model (Salton et al., 1975), each sentence s_i can be represented by a vector $\vec{s}_i$ with each component being the weight of a term in s_i. The weight associated with a term w is calculated by $tf(w) * isf(w)$, where $tf(w)$ is the frequency of the term w in sentence s_i and $isf(w)$

is the inverse sentence frequency[4] of term w: $isf(w) = 1 + \log(N / n_w)$, where N is the total number of sentences in S and n_w is the number of sentences containing w.The similarity score between two sentences is computed using the inner product of the corresponding sentence vectors, as follows:

$$sim(s_i, s_j) = \frac{\vec{s}_i \cdot \vec{s}_j}{\| \vec{s}_i \| \cdot \| \vec{s}_j \|}$$

Taking each sentence as a vertex, and the similarity score as the weight of the edge between two sentences, a sentence graph is constructed. The graph is fully connected and undirected because the similarity score is symmetric.

The sentence graph can be modeled by an adjacency matrix $\mathbf{M}$, in which each element corresponds to the weight of an edge in the graph. Thus $\mathbf{M} = [M_{ij}]_{N \times N}$ is defined as:

$$M_{ij} = \begin{cases} \dfrac{\vec{s}_i \cdot \vec{s}_j}{\| \vec{s}_i \| \cdot \| \vec{s}_j \|}, & if\ i \neq j \\ 0, & otherwise \end{cases}$$

We normalize the row sum of matrix $\mathbf{M}$ in order to assure it is a stochastic matrix such that the PageRank iteration algorithm is applicable. The normalized matrix is:

$$\tilde{M}_{ij} = \begin{cases} M_{ij} \Big/ \sum_{j=1}^{N} M_{ij}, & if \sum_{j=1}^{N} M_{ij} \neq 0 \\ 0, & otherwise \end{cases}.$$

Using the normalized adjacency matrix, the salience score of a sentence s_i is computed in an iterative manner:

$$score(s_i) = d \cdot \sum_{j=1}^{N} score(s_j) \cdot \tilde{M}_{ji} + \frac{(1-d)}{N}$$

where d is a damping factor that is typically between 0.8 and 0.9 (Page et al., 1998).

If we use a column vector p to denote the salience scores of all the sentences in S, the above equation can be written in a matrix form as follows:

$$p = [d \cdot \mathbf{M}^T + (1-d) \cdot \mathbf{U}] \cdot p$$

where $\mathbf{U}$ is a square matrix with all elements being equal to $1/N$. The component $(1-d) \cdot \mathbf{U}$ can be considered as a smoothing term which adds a small probability for a random walker to jump from the current vertex to any vertex in the graph. This guarantees that the stochastic transition matrix for iteration is irreducible and aperiodic. Therefore the iteration can converge to a stable state.

In our implementation, the damping factor d is set to 0.85 as in the PageRank algorithm (Page et al., 1998). The column vector $\boldsymbol{p}$ is initialized with random values between 0 and 1. After the algorithm converges, each component in the column vector $\boldsymbol{p}$ corresponds to the salience score of the corresponding sentence. This score is combined with the GO relevance score to rank sentences.

3.3 GO Relevance Scoring

Up to this point, our system considers only gene-independent features, in both sentence filtering and PageRank-based sentence scoring. These features are universal across different genes. However, each gene is unique because of its own functional and structural properties. Thus we seek to include gene-specific features in this next step.

The GO annotations provide one kind of gene-specific information and have been shown to be useful for selecting GeneRIF candidates (Lu et al., 2006). A gene's GO annotations include descriptions in three aspects: molecular function; biological process; and cellular component. For example, the human gene AANAT (gene ID 15 in Entrez Gene) is annotated with the GO terms in Table 4.

GO ID	*GO term*
GO:0004059	aralkylamine N-acetyltransferase activity
GO:0007623	circadian rhythm
GO:0008152	metabolic process
GO:0008415	acyltransferase activity
GO:0016740	transferase activity

Table 4. GO terms for gene AANAT

The GO relevance score is computed as follows: first, the GO terms and the sentences are both stemmed and stopwords are removed. For example, the GO terms in Table 4 are processed into a set of stemmed words: *aralkylamin, N, acetyltransferas, activ, circadian, rhythm, metabol, process, acyltransferas* and *transferas*.

[4] *Isf* is equivalent to *idf* if we view each sentence as a document.

Second, the total number of occurrence of the GO terms appearing in a sentence is counted. Finally, the GO relevance score is computed as the ratio of the total occurrence to the sentence length. The entire process can be illustrated by the following pseudo codes:

```
1 tokenize and stem the GO terms;
2 tokenize and stem all the sentences, remove stop words;
3 for each sentence $s_i$ , $i = 1, ..., N$
    $GOScore(s_i) = 0$
    for each word $w$ in $s_i$
        if $w$ in the GO term set
            $GOScore(s_i)$ ++
        end if
    end for
    $GOScore(s_i) = GOScore(s_i) / length(s_i)$
end for
```

where $length(s_i)$ is the number of distinct non-stop words in s_i . For each sentence s_i , the GO relevance score is combined with the PageRank score to get the overall score (α is a weight parameter between 0 and 1; see Section 4.2 for discussion):

$$score(s_i) = \alpha \cdot PRScore(s_i) + (1 - \alpha) \cdot GOScore(s_i) .$$

3.4 Redundancy Removal

A good summary contains as much diverse information as possible for a gene, while with as little redundancy as possible. For many well-studied genes, there are thousands of relevant papers and much information is redundant. Hence it is necessary to remove redundant sentences before producing a final summary.

We adopt the diversity penalty method (Zhang et al., 2005; Wan and Xiao, 2007) for redundancy removal. The idea is to penalize the candidate sentences according to their similarity to the ones already selected. The process is as follows:

(1) Initialize two sets, $A = \phi$, $B = \{s_i \mid i = 1, 2, ..., K\}$ containing all the extracted sentences;

(2) Sort the sentences in B by their scores in descending order;

(3) Suppose s_i is the top ranked sentence in B , move it from B to A . Then we penalize the remaining sentences in B as follows:

For each sentence s_j in B , $j \neq i$

$$Score(s_j) = Score(s_j) - \omega \cdot sim(s_j, s_i) \cdot Score(s_i)$$

where $\omega > 0$ is the penalty degree factor, $sim(s_j, s_i)$ is the similarity between s_i and s_j .

(4) Repeat steps 2 and 3 until enough sentences have been selected.

4 Results and Discussion

4.1 Evaluation Metrics

Unlike the newswire summarization, there are no gold-standard test collections available for evaluating gene summarization systems. The two previous studies mentioned in Section 2 both conducted *extrinsic* evaluations by asking human experts to rate system outputs. Although it is important to collect direct feedback from the users, involving human experts makes it difficult to compare different summarization systems and to conduct large-scale evaluations (both studies evaluated nothing but a small number of genes). In contrast, we evaluated our system intrinsically on a much larger dataset consisting of 7,294 human genes, each with a preexisting handwritten summary downloaded from the NCBI's FTP site[5].

The handwritten summaries were used as reference summaries (i.e. a gold standard) to compare with the automatically generated summaries. Although the length of reference summaries varies, the majority of these summaries contain 80 to 120 words. To produce a summary of similar length, we decided to select five sentences consisting of about 100 words.

For the intrinsic evaluation of a large number of summaries, we made use of the ROUGE metrics that has been widely used in automatic evaluation of summarization systems (Lin and Hovy, 2003; Hickl et al., 2007). It provides a set of evaluation metrics to measure the quality of a summary by counting overlapping units such as n-grams or word sequences between the generated summary and its reference summary.

[5] ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/

We computed three ROUGE measures for each summary, namely ROUGE-1 (unigram based), ROUGE-2 (bigram based) and ROUGE-SU4 (skip-bigram and unigram) (Lin and Hovy, 2003). Among them, ROUGE-1 has been shown to agree most with human judgments (Lin and Hovy, 2003). However, as biomedical concepts usually contain more than one word (e.g. transcription factor), ROUGE-2 and ROUGE-SU4 scores are also important for assessing gene summaries.

4.2 Determining parameters for best performance

The two important parameters in our system – the linear coefficient α for the combination of PageRank and GO scores and the diversity penalty degree factor ω in redundancy removal – are investigated in detail on a collection of 100 randomly selected genes. First, by setting α to values from 0 to 1 with an increment of 0.1 while holding ω steady at 0.7, we observed the highest ROUGE-1score when α was 0.8 (Figure 2). This suggests that the two scores (i.e. PageRank and GO score) complement to each other and that the PageRank score plays a more dominating role in the summed score. Next, we varied ω gradually from 0 to 5 with an increment of 0.25 while holding α steady at 0.75.The highest ROUGE-1 score was achieved when ω was 1.3 (Figure 3). For ROURE-2, the best performance was obtained when α was 0.7 and ω was 0.5. In order to balance ROUGE-1 and ROUGE-2 scores, we set α to 0.75 and ω to 0.7 for the remaining experiments.

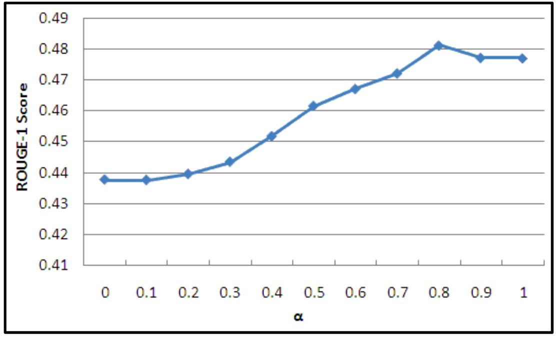

Figure 2. The blue line represents the changes in ROUGE-1 scores with different values of α while ω is held at 0.7.

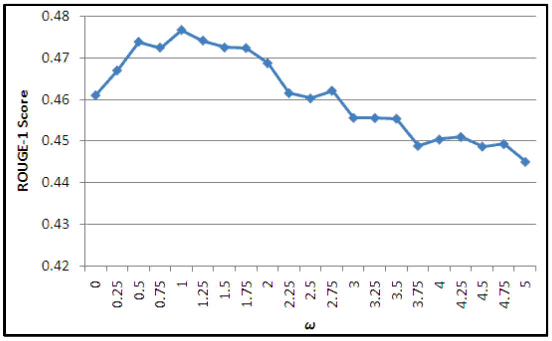

Figure 3. The blue line represents the changes in ROUGE-1 scores with different values of ω while α is held at 0.75.

4.3 Comparison with other methods

Because there are no publicly available gene summarization systems, we compared our system with two baseline methods. The first is a well known publicly available summarizer - MEAD (Radev et al., 2004). We adopted the latest version of MEAD 3.11 and used the default setting in MEAD that extracts sentences according to three features: centroid, position and length. The second baseline extracts different sentences randomly from abstracts. Comparison results are shown in the following table:

System	*ROUGE-1*	*ROUGE-2*	*ROUGE-SU4*
Our System	0.4725	0.1247	0.1828
MEAD	0.3890	0.0961	0.1449
Random	0.3434	0.0577	0.1091

Table 5. Systems comparison on 7,294 genes.

As shown in Table 5, our system significantly outperformed the two baseline systems in all three ROUGE measures. Furthermore, larger performance gains are observed in ROUGE-2 and ROUGE-SU4 than in ROUGE-1. This is because many background words (e.g. *gene*, *protein* and *enzyme*) also appeared frequently as unigrams in randomly selected summaries.

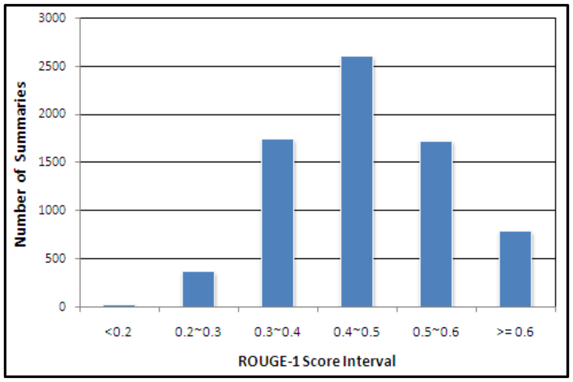

Figure 4. ROUGE-1 score distribution

In Figure 4, we show that the majority of the summaries have a ROUGE-1 score greater than 0.4. Our further analysis revealed that almost half summaries with a low score (smaller than 0.3) either lacked sufficient relevant abstracts, or the reference summary was too short or too long. In either case, only few overlapping words can be found when comparing the generated gene summary with the reference. The statistics for low ROUGE-1 score are listed in Table 6. We also note that almost half of the summaries that have low ROUGE-1 scores were due to other causes: mostly, machine generated summaries differ from human summaries in that they describe different functional aspects of the same gene product. Take the gene TOP2A (ID: 7153) for example. While both summaries (handwritten and machine generated) focus on its encoded protein *DNA topoisomerase*, the handwritten summary describes the chromosome location of the gene whereas our algorithm selects statements about its gene expression when treated with a chemotherapy agent. We plan to investigate such differences further in our future work.

Causes for Low Score	*Number of genes*
Few (≤10) related abstracts	106
Short reference summary (< 40 words)	27
Long reference summary (> 150 words)	76
Other	198
Total	407

Table 6. Statistics for low ROUGE-1 scores (<0.3)

4.4 Results on various summary length

Figure 5 shows the variations of ROUGE scores as the summary length increases. At all lengths and for both ROUGE-1 and ROUGE-2 measures, our proposed method performed better than the two baseline methods. By investigating the scores of different summary lengths, it can be seen that the advantage of our method is greater when the summary is short. This is of great importance for a summarization system as ordinary users typically prefer short content for summaries.

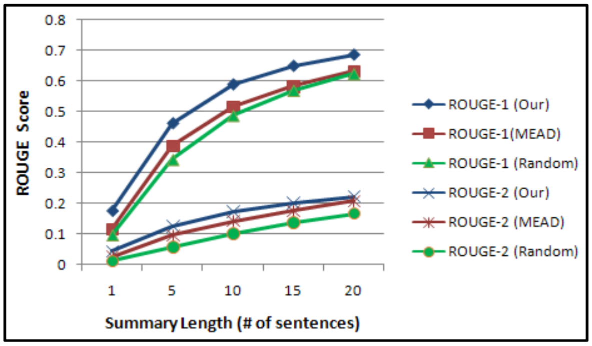

Figure 5. Score variation for different summary length

5 Conclusions and Future Work

In this paper we have presented a system for generating gene summaries by automatically finding extract-worthy sentences from the biomedical literature. By using the state-of-the-art summarization techniques and incorporating gene specific annotations, our system is able to generate gene summaries more accurately than the baseline methods. Note that we only evaluated our system for human genes in this work. More summaries are available for human genes than other organisms, but our method is organism-independent and can be applied to any other species.

This research has implications for real-world applications such as assisting manual database curation or updating existing gene records. The ROUGE scores in our evaluation show comparable performance to those in the newswire summarization (Hickl et al., 2007). Nonetheless, there are further steps necessary before making our system output readily usable by human curators. For instance, human curators are generally in favor of sentences presented in a coherent order. Thus, information-ordering algorithms in multi-document summarization need to be investigated. We also plan to study the guidelines and scope of the curation process, which may provide additional important heuristics to further refine our system output.

Acknowledgments

The work is supported by NSFC project No. 60803075, Chinese 973 Project No. 2007CB311003. ZL is supported by the Intramural Program of the National Institutes of Health. The authors are grateful to W. John Wilbur and G. Craig Murray for their help on the early version of this manuscript.

References

W. A. Baumgartner, B. K. Cohen, L. M. Fox, G. Acquaah-Mensah, L. Hunter. 2007. Manual Curation Is Not Sufficient for Annotation of Genomic Databases. Bioinformatics, Vol. 23, No. 13. (July 2007), pp. i41-48.

G. Bhalotia, P. I. Nakov, A. S. Schwartz and M. A. Hearst, BioText Team Report for the TREC 2003 Genomics Track. In Proceedings of TREC 2003.

B. Boguraev and C. Kennedy. 1997. Salience-based Content Characterization of Text Documents. In Proceedings of Workshop on Intelligent Scalable Text Summarization (ACL97/EACL97), pp. 2-9.

J. Carbonell and J. Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In ACM SIGIR, pages 335–336, August.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. Journal of the ACM (JACM) archive Volume 16, Issue 2 (April 1969) Pages: 264 – 285.

G. Erkan and D. R. Radev. 2004. LexPageRank: Prestige in Multi-Document Text Summarization. In Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain.

S. Gupta, A.Nenkova and D.Jurafsky. 2007. Measuring Importance and Query Relevance in Topic-focused Multi-document Summarization. Proceedings of ACL 2007 short papers, Prague, Czech Republic.

W. Hersh and R. T. Bhupatiraju. 2003. TREC Genomics track Overview. In Proceedings of TheTwelfth Text REtrieval Conference, 2003.

A. Hickl, K. Roberts and F. Lacatusu. 2007. LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization.

R. Jelier, M. Schuemie, C. Eijk, M. Weeber, E. Mulligen, B. Schijvenaars, B. Mons, J. Kors. Searching for geneRIFs: Concept-based Query Expansion and Bayes Classification. In Proceedings of TREC 2003.

C. Lin and E. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In Proceedings of the COLING Conference.

C. Lin and E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In HLT-NAACL, pages 71–78.

X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai and B. Schatz. 2006. Automatically Generating Gene Summaries from Biomedical Literature. Proceedings of the Pacific Symposium on Biocomputing 2006.

X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai and B. Schatz. 2007. Generating Gene Summaries from Biomedical Literature: A Study of Semi-Structured Summarization. Information Processing and Management 43, 2007, 1777-1791.

Z. Lu, K. B. Cohen and L. Hunter. 2006. Finding GeneRIFs via Gene Ontology Annotations. Pac Symp-Biocomput. 2006:52-63.

C. Manning and H. Schütze. 1999. Foundations of Statistical Natural Language Processing. Chapter 5, MIT Press. Cambridge, MA: May 1999.

K. R. McKeown and D. R. Radev. 1995. Generating Summaries of Multiple News Articles. In Proceedings, ACM Conference on Research and Development in Information Retrieval SIGIR'95, pages 74–82.

R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, July 2004.

M. Newman. 2003. The Structure and Function of Complex Networks. SIAM Review 45.167–256 (2003).

L. Page, S. Brin, R. Motwani and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, Stanford, CA, 1998.

D. R. Radev, H. Jing, M. Stys and D. Tam. 2004. Centroid-based Summarization of Multiple Documents. Information Processing and Management, 40:919–938.

G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, nr.11, pages 613–620.

S. Teufel and M. Moens. 1997. Sentence Extraction as a Classification Task. Workshop 'Intelligent and scalable Text summarization', ACL/EACL 1997.

X. Wan and J. Xiao. 2007. Towards a Unified Approach Based on Affinity Graph to Various Multi-document Summarizations. ECDL 2007: 297-308.

J. Yang, A. M. Cohen, W. Hersh. Automatic Summarization of Mouse Gene Information by Clustering and Sentence Extraction from MEDLINE Abstracts. AMIA 2007 Annual Meeting. Nov. 2007 Chicago, IL.

J. Yang, A. M. Cohen, W. Hersh. 2008. Evaluation of a Gene Information Summarization System by Users During the Analysis Process of Microarray Datasets. In BMC Bioinformatics 2009 10(Suppl 2):S5.

B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, W. Ma. 2005. Improving Web Search Results Using Affinity Graph. The 28th Annual International ACM SIGIR Conference (SIGIR'2005), August 2005.