Surveyor: A System for Generating Coherent Survey Articles for Scientific Topics

Rahul Jha and Reed Coke
Department of EECS
University of Michigan
Ann Arbor, MI, 48109

Abstract
We investigate the task of generating coherent survey articles for scientific topics. We introduce an extractive summarization algorithm that combines a content model with a discourse model to generate coherent and readable summaries of scientific topics using text from scientific articles relevant to the topic. Human evaluation on 15 topics in computational linguistics shows that our system produces significantly more coherent summaries than previous systems. Specifically, our system improves the ratings for coherence by 36% in human evaluation compared to C-Lexrank, a state of the art system for scientific article summarization.

Introduction
This paper is about generating coherent summaries of scientific topics. Given a set of input papers that are relevant to a specific topic such as question answering, our system called Surveyor extracts and organizes text segments from these papers into a coherent and readable survey of the topic. There are many applications for automated surveys thus generated. Human surveys do not exist for all topics and quickly become outdated in rapidly growing fields like computer science. Therefore, an automated system for this task can be very useful for new graduate students and cross-disciplinary researchers who need to quickly familiarize themselves with a new topic.

Our work builds on previous work on summarization of scientific literature (Mohammad et al. 2009; Qazvinian and Radev 2008). Prior systems for generating survey articles for scientific topics such as C-Lexrank have focused on building informative summaries but no attempt has been made to ensure the coherence and readability of the output summaries. Surveyor on the other hand focuses on generating survey articles that contain well defined subtopics presented in a coherent order. Figure 1 shows part of the output of Surveyor for the topic of question answering.

Our experimental results on a corpus of computational linguistics topics show that Surveyor produces survey articles that are substantially more coherent and readable compared to previous work. The main contributions of this paper are:

• We propose a summarization algorithm that combines a content model and a discourse model in a modular way to build coherent summaries.

• We introduce the notion of Minimum Independent Discourse Contexts as a way of flexibly modeling discourse relationships in a summarization system.

• We conducted several experiments for evaluating coherence and informativeness of Surveyor on a dataset of 15 topics in computational linguistics with 297 articles and 30 human-written gold summaries (2 per topic). All data used for our experiments is available at http://clair.si.umich.edu/corpora/surveyor_aaaai_15.tgz.

We first give an overview of our summarization approach. This is followed by details about our experimental setup and a discussion of results. Finally, we summarize the related work and conclude the paper with pointers for future work.
Overview of Summarization Approach

We first describe the two main components of our system and then describe our summarization algorithm that is built on top of them.

Content Model Given a set of research papers relevant to a scientific topic, each of them focusing on a specific aspect of the problem. For example, a paper on supervised word sense disambiguation might describe the background on word sense disambiguation followed by a review of supervised methods for the problem. Similarly, a paper on unsupervised word sense disambiguation may give some general overview of the field, then briefly describe supervised approaches followed by a more detailed overview of unsupervised methods. We capture these subtopics in the input documents and their transitions using a Hidden Markov Model (HMM) where the states of the HMM correspond to subtopics. Given the set of \( k \) subtopics \( S = (s_1 \cdots s_k) \), the state transitions of the HMM are defined as:

\[
p(s_j | s_i) = \frac{\text{Count}(s_i, s_j) + \delta}{\text{Count}(s_i) + \delta \cdot m}
\]

Where \( \text{Count}(s_i, s_j) \) is the number of times a sentence from subtopic \( s_j \) appears immediately after a sentence from subtopic \( s_i \) in the input document collection and \( \text{Count}(s_i) \) is the total number of times the subtopic \( s_i \) appears in the input document set. \( \delta \) is a smoothing parameter and \( m \) is the number of sentences in \( s_i \).

To initialize the states of the HMM, we use a network based clustering approach. We build a lexical network where the sentences represent the nodes of the network and the edge weights are the \( \text{tf} \times \text{idf} \) similarity between each pair of sentences \(^1\). Given this lexical network, we use the Louvain clustering method (De Meo et al. 2011) to partition the lexical network into clusters. Each cluster in the network is then initialized to a sub-topic. Louvain is a hierarchical clustering algorithm that does not need the number of output clusters as a parameter. The HMM is then learned through Viterbi decoding. Our HMM model is similar to (Barzilay and Lee 2004), but we take the novel step of using the transition matrix to guide the summarization output, as described below.

Table 1: A partial table of transition probabilities between three subtopics for word sense disambiguation. The probabilities do not add up to 1 because the table only shows a few states from a larger transition matrix.

<table>
<thead>
<tr>
<th></th>
<th>subtopic 1</th>
<th>subtopic 2</th>
<th>subtopic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>start</td>
<td>0.35</td>
<td>0.50</td>
<td>0</td>
</tr>
<tr>
<td>subtopic 1</td>
<td>0.49</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>subtopic 2</td>
<td>0.24</td>
<td>0.41</td>
<td>0.02</td>
</tr>
<tr>
<td>subtopic 3</td>
<td>0.25</td>
<td>0.23</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Figure 2 shows sentences from three of the subtopics learned for the topic of word sense disambiguation. In a coherent summary, subtopic 2 containing background sentences should appear before subtopic 1 that contains details about a specific method. We use the transition matrix of the learned HMM to model these subtopic transitions in the original documents and use it to guide the summarizer output. As an example, a partial table of transition probabilities learned for the subtopics in Figure 2 is shown in Table 1, where \( \text{start} \) is a pseudo-state representing the beginning of the document. The highest outgoing probability from \( \text{start} \) is to subtopic 2, which allows the summarizer to include background information about the topic at the beginning followed by sentences from more specific subtopics represented by subtopic 1 and subtopic 3.

Discourse Model A common problem with extractive summaries is that the sentences used from the original input documents may not be understandable when pulled out of their original context. To avoid such problems, we introduce the idea of Minimum Independent Discourse Contexts (MIDC).

Definition. Given a text segment \( T \) containing \( n \) sentences \( (s_1 \cdots s_n) \), the minimum independent discourse context (midc) of a sentence \( s_i \) is defined as the minimum set of \( j \) sentences \( \text{midc}(s_i) = (s_{i-j} \cdots s_i) \) such that given \( \text{midc}(s_i), s_i \) can be interpreted independently of the other sentences in \( T \).

Figure 3 shows how this idea works in practice. Sentences \( s_1 \) and \( s_4 \) can be included in a summary without requiring additional context sentences. Sentences \( s_2 \), \( s_3 \) and \( s_4 \) on the

---

\(^1\)The idfs are computed over the entire input corpus.
Opinion words are words that convey positive or negative polarities.

They are critical for opinion mining (Pang et al., 2002; Turney, 2002; Hu and Liu, 2004; Wilson et al., 2004; Popescu and Eitzioni, 2005; Gamon et al., 2005; Ku et al., 2006; Brecc et al., 2007; Kobayashi et al., 2007; Ding et al., 2008; Titov and McDonald, 2008; Pang and Lee, 2008; Lu et al., 2009).

The key difficulty in finding such words is that opinions expressed by many of them are domain or context dependent.

Several researchers have studied the problem of finding opinion words (Liu, 2010).

The approaches can be grouped into corpus-based approaches (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Kanayama and Nasukawa, 2006; Qiu et al., 2009) and dictionary-based approaches (Hu and Liu, 2004; Kim and Hovy, 2004; Kamps et al., 2004; Esuli and Sebastiani, 2005; Takamura et al., 2005; Andreewksaia and Bergler, 2006; Dragut et al., 2010).

The approaches can be grouped into corpus-based approaches (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Kanayama and Nasukawa, 2006; Qiu et al., 2009) and dictionary-based approaches (Hu and Liu, 2004; Kim and Hovy, 2004; Kamps et al., 2004; Esuli and Sebastiani, 2005; Takamura et al., 2005; Andreewksaia and Bergler, 2006; Dragut et al., 2010).

Table 2: Discourse rules used to create minimum independent discourse contexts.

<table>
<thead>
<tr>
<th>Discourse relationship</th>
<th>Dependency rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreference</td>
<td>Add a dependency between $s_i$ and $s_j$ if they belong to a coreference chain.</td>
</tr>
<tr>
<td>Discourse Transition</td>
<td>Add a dependency between $s_{i-1}$ and $s_i$ if $s_i$ contains an explicit discourse marker.</td>
</tr>
<tr>
<td>Entity Transition</td>
<td>Add a dependency between $s_i$ and $s_j$ if they both share a prominent entity.</td>
</tr>
</tbody>
</table>

Figure 3: A paragraph from an input paper on the topic of opinion mining along with the $midc$ for each sentence on the right.

Figure 4: Summarization Algorithm.

The algorithm accepts a set of input documents $docs$ and a maximum summary length $maxlen$. It first learns the subtopics and their transition matrix by running HMM on the input document set. After initializing the first subtopic to the pseudo-subtopic $start$, it iteratively picks the next subtopic by using the HMM transition matrix. Given each subtopic, it runs a salience algorithm on all the sentences of the subtopic to find the most central sentence of the subtopic. In the current implementation, this is done using Lexrank (Erkan and Radev 2004). Given the subtopic’s most central sentence, it calculates the $midc$ for this sentence and if the $midc$ is valid, it is added to the output summary. An $midc$ can be invalid if it exceeds a maximum threshold number of sentences. The $midc$ is then removed from the subtopic so it will not be picked if we visit this subtopic again. This procedure continues until we obtain a summary of the desired length. Important subtopics in the input can get more than one $midc$ in the summary because the transition matrix contains high probabilities for transitioning to these subtopics.

Table 2: Discourse rules used to create minimum independent discourse contexts.
Experimental Setup

The main research questions that we want to answer using our experiments are:

1. Are the summaries created using Surveyor more coherent than previous state-of-the-art methods for survey article generation?

2. What are the individual contributions of the content model and the discourse model?

3. How does Surveyor compare against state-of-the-art systems for coherent news summarization applied to the survey generation problem?

For research question 1, we compare our system with C-Lexrank (Mohammad et al. 2009), a state-of-the-art system for survey generation. For research question 2, we measure the effects of HMM and MIDC models in isolation on the quality of output summaries. For research question 3, we compare our system with G-Flow (Christensen et al. 2013), a state-of-the-art system for coherent summarization of news articles. We now describe the data used in our experiments.

Data

We used the ACL Anthology Network (AAN) (Radev et al. 2013) as a corpus for our experiments and selected 15 established topics in computational linguistics for our evaluation. The input documents used for summarization of a research topic should be research papers that describe the most relevant research in the topic. Since the focus of this paper is on summarization, we used an oracle method for selecting the initial set of papers for each topic. We collected at least three human-written surveys on each topic. The bibliographies of all the surveys were processed using Parscit (Luong, Nguyen, and Kan 2010) and any document that appeared in the bibliography of more than one survey was added to the initial document set $D_i$.

An ideal survey article on the topic should describe the research represented by $D_i$. These sentences are actually found in papers that cite papers in $D_i$ and thus describe their contributions. Therefore to create the final document set $D_f$, we collect all the papers in AAN that cite the papers in $D_i$. The citing documents are then ordered based on the number of papers in $D_i$ that they cite and the top $n$ documents are added to $D_f$. The text input for the summarization system is extracted from $D_f$. For our current experiments, the value of $n$ is set to 20.

For the task of survey article generation, the most relevant text is found in the introduction sections of $D_f$ since this is where researchers describe the prior work done by subsets of papers in $D_i$. Therefore, we extract the sentences in the introductions of each of the papers in $D_f$ as the text input for our summarizer. Table 3 shows the set of 15 topics and size of summarizer input for each topic.

Experiments

Coherence Evaluation with C-Lexrank  For coherence evaluation, we generated fixed length 2000 character summaries using both C-Lexrank and Surveyor. Six assessors with background in computational linguistics manually evaluated pairs of output summaries. Given two summaries, the assessors were asked to mark which summary they preferred, or mark “indifferent” if they could not choose one against the other. The presentation of summary pairs to assessors as well as the order of summaries in each pair was randomized. Thus, the assessors had no way of telling which systems produced the pair of summaries they saw.

Compared to C-Lexrank, the assessors preferred a summary generated by Surveyor 67% of the time and were indifferent 20% of the time (Table 4).

Table 4: Overall summary preference for Surveyor compared to C-Lexrank.

<table>
<thead>
<tr>
<th>Surveyor</th>
<th>Indifferent</th>
<th>C-Lexrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>67%</td>
<td>20%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Additionally, the assessors were asked to rate each summary based on the standard DUC quality questions 4. The DUC quality questions are a standard benchmark used for evaluating summaries on the aspects of overall coherence, avoiding useless text, avoiding repetitive information, avoiding bad referents and avoiding overly explicit referents. For each of the questions, the assessors can assign a score from 1 to 5 with higher being better.

As shown in Table 5, the assessors also assigned much higher scores to summaries generated by Surveyor on an average compared to C-Lexrank on all the DUC quality questions 5. On the metric of coherence, the scores for Surveyor

Table 5: Average DUC quality scores for Surveyor and C-Lexrank.

<table>
<thead>
<tr>
<th>Question</th>
<th>Surveyor</th>
<th>C-Lexrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.9</td>
<td>2.7</td>
</tr>
</tbody>
</table>

5On average, we found only 33% of the documents in $D_i$ to be in AAN. Since the citation network for AAN contains only citations within AAN documents, we implemented a record matching algorithm to find all the papers in AAN that cite any arbitrary document outside AAN.


5DUC quality responses represent a Likert-type scale. The use of parametric statistics such as mean for such data has been debated, but there are several recent arguments for its validity (Norman 2010).
Table 5: Average scores on the DUC quality questions for C-Lexrank and different Surveyor variants along with standard error.

<table>
<thead>
<tr>
<th>Quality Question</th>
<th>C-Lexrank</th>
<th>Surveyor</th>
<th>Surveyor HMM Only</th>
<th>Surveyor MIDC Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>coherence</td>
<td>2.72 ± 0.16</td>
<td>3.70 ± 0.22</td>
<td>2.57 ± 0.20</td>
<td>3.07 ± 0.36</td>
</tr>
<tr>
<td>avoid useless text</td>
<td>3.20 ± 0.15</td>
<td>3.90 ± 0.15</td>
<td>3.17 ± 0.19</td>
<td>3.33 ± 0.30</td>
</tr>
<tr>
<td>avoid repetition</td>
<td>4.07 ± 0.11</td>
<td>4.23 ± 0.14</td>
<td>3.97 ± 0.19</td>
<td>4.40 ± 0.19</td>
</tr>
<tr>
<td>avoid bad referents</td>
<td>3.43 ± 0.16</td>
<td>4.17 ± 0.14</td>
<td>3.60 ± 0.18</td>
<td>3.47 ± 0.27</td>
</tr>
<tr>
<td>avoid overly explicit referents</td>
<td>4.23 ± 0.12</td>
<td>4.47 ± 0.11</td>
<td>4.30 ± 0.19</td>
<td>4.53 ± 0.22</td>
</tr>
</tbody>
</table>

Table 6: Overall summary preference for the two Surveyor variants compared to C-Lexrank.

<table>
<thead>
<tr>
<th></th>
<th>Surveyor HMM Only</th>
<th>Indifferent</th>
<th>C-Lexrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveyor HMM Only</td>
<td>53%</td>
<td>27%</td>
<td>20%</td>
</tr>
<tr>
<td>Surveyor MIDC Only</td>
<td>33%</td>
<td>27%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Informativeness Evaluation We use ROUGE (Lin 2004b) for informativeness evaluation. ROUGE is a standard evaluation metric for automatic evaluation of summaries that uses n-gram co-occurrences between automated summaries and human generated reference summaries to score the automated summaries. ROUGE has been shown to correlate well with human evaluations (Lin 2004a).

For ROUGE evaluation, we asked two assessors to generate 2000 character long summaries using the input for each topic. We then did ROUGE evaluation of the summaries generated using C-Lexrank and Surveyor against these gold summaries. The average ROUGE-1 and ROUGE-2 scores are summarized below in Table 7. The improvement in ROUGE scores of Surveyor over C-Lexrank is statistically significant with $p < 0.05$. Thus Surveyor, in addition to producing more coherent summaries, also produces summaries that are more informative given the same input text.

Table 7: Average Rouge scores for each of the systems for 15 topics evaluated against two reference summaries per topic.

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Lexrank</td>
<td>0.40</td>
<td>0.05</td>
</tr>
<tr>
<td>Surveyor</td>
<td>0.44</td>
<td>0.19</td>
</tr>
<tr>
<td>Surveyor HMM Only</td>
<td>0.42</td>
<td>0.13</td>
</tr>
<tr>
<td>Surveyor MIDC only</td>
<td>0.42</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Evaluation with G-FLOW G-FLOW (Christensen et al. 2013) is a recent state of the art system for generating co-
herent summaries that has been evaluated on newswire data. We compared Surveyor with G-FLOW by running the im-
plementation of G-FLOW obtained from the original au-
thors on our evaluation data. The coherence evaluation with
G-FLOW was done in the same way as for C-Lexrank except
the output summary length for both systems was limited to
1000 characters. This is because the optimization procedure
implemented in G-FLOW becomes intractable for output of
2000 characters.

In the coherence evaluation, assessors preferred Survey-

or 47% of the time compared to 40% of the time for
ROUGE-1 and ROUGE-2 improvements of Surveyor over
LOW Surveyor compared with G-FLOW.

In informativeness evaluation with ROUGE, the 1000
characters generated by Surveyor got an average ROUGE-1 score of 0.41 compared to a score of 0.36
obtained by G-FLOW. The ROUGE-2 score of Surveyor
was 0.13 compared to 0.07 for G-FLOW. p-values for the
ROUGE-1 and ROUGE-2 improvements of Surveyor over
G-FLOW are 0.12 and 0.11 respectively. These results in-
dicate that Surveyor does slightly better than G-FLOW in
terms of coherence evaluation while also producing in-
formative summaries. This suggests that the HMM based
content model does a better job of modeling the flow of
subtopics in scientific articles compared to G-FLOW which

does not include such a component.

Related Work

Multi-document summarization of scientific articles has
been studied by Nanba, Kando, and Okumura (2004). Moh-
hammad et al. (2009) compared several algorithms for gen-
erating automated surveys of scientific topics. Jha, Abu-
Jbara, and Radev (2013) implemented a system that can
summarize a topic starting from a query as input. However,
none of these papers focused on evaluating the coherence of
resulting summaries. In the medical domain, several sum-
marization systems have been proposed that take advantage
of the rich ontological data available for medical concepts
(Elhadad and McKeown 2001; Kan, McKeown, and Kl-
vans 2001; Yoo, Hu, and Song 2006). A different stream
of research has looked at summarizing scientific research
using the metaphor of maps (Fried and Kobourov 2013;
Shahaf, Guestrin, and Horvitz 2012). For the work on single
document summarization for scientific literature, we refer
the readers to the review in Nenkova and McKeown (2011).

Ordering the sentences in summarization output for im-
proving readability has been studied by Barzilay and McK-
eown (2005) and Bollegala, Okazaki, and Ishizuka (2010).
Automatic metrics for estimating coherence for summar-
ization evaluation have also been studied (Lapata 2005;
Lin et al. 2012). More recently, Christensen et al. (2013)
presented an algorithm called G-FLOW for joint sentence
selection and ordering for news summarization.

Barzilay and Lee (2004) and Fung and Ngai (2006) have
presented HMM based content models that use the HMM
topics as features in a supervised summarization system to
produce informative summaries. LDA (Latent Dirichlet Al-
location) based content models for summarizing documents
(Daumé and Marcu 2006; Haghighi and Vanderwende 2009)
have also been explored, but they focus on maximizing in-
formativeness instead of coherence.

Conclusion and Future Work

In this paper, we present Surveyor, a system for generating
coherent surveys of scientific articles. We describe our al-
gorithm and present experimental results on a corpus of 15
topics in computational linguistics. Our results show that our
system leads to more coherent summaries than C-Lexrank,
a state-of-the-art system for survey article generation and
G-FLOW, a state-of-the-art system for coherent summariza-
tion. In particular, in human evaluation for coherence, Sur-
veyor outperforms the performance of C-Lexrank by 36%
and outperforms the performance of G-FLOW by 4%.

This work suggests several possible future directions for
research. The first is developing more sophisticated content
models that better capture the distribution of topics in sci-
entific documents across genres. The second is building a
corpus of discourse relationships between sentences in sci-
centific documents as well as improving the algorithm for cre-
ating minimum independent discourse context. Finally, au-
tomatic sentence compression, fusion and rewriting strate-
gies can be applied to sentences of the output summary to
remove irrelevant text segments and improve the informa-
tiveness of the summaries.

Acknowledgments

We thank the anonymous reviewers for their feedback on
earlier versions of this paper. Also, thanks to Janara Chris-
tensen for providing results of running G-FLOW on our
data. Finally, thanks to Inderjeet Mani, Zornitsa Kozareva,
Or Biran, Ben King, Veronica Perez Rosas, Charlie Welch,
Shibamouli Lahiri, Max Kaufmann and Shiwali Mohamed for
several useful discussions and help with evaluations.