

A System for Summarizing Scientific Topics Starting from Keywords

Rahul Jha

Department of EECS
University of Michigan
Ann Arbor, MI, USA
rahuljha@umich.edu

Amjad Abu-Jbara

Department of EECS
University of Michigan
Ann Arbor, MI, USA
amjbara@umich.edu

Dragomir Radev

Department of EECS and
School of Information
University of Michigan
Ann Arbor, MI, USA
radev@umich.edu

Abstract

In this paper, we investigate the problem of automatic generation of scientific surveys starting from keywords provided by a user. We present a system that can take a topic query as input and generate a survey of the topic by first selecting a set of relevant documents, and then selecting relevant sentences from those documents. We discuss the issues of robust evaluation of such systems and describe an evaluation corpus we generated by manually extracting factoids, or information units, from 47 gold standard documents (surveys and tutorials) on seven topics in Natural Language Processing. We have manually annotated 2,625 sentences with these factoids (around 375 sentences per topic) to build an evaluation corpus for this task. We present evaluation results for the performance of our system using this annotated data.

1 Introduction

The rise of the number of publications in all scientific fields is making it more and more difficult to get quickly acquainted with the new developments in a new area. One way to wade through this huge amount of scholarly information is to consult topical surveys written by experts in an area. For example, for machine translation, one might read (Lopez, 2008)¹. Such surveys can be very helpful when available, but unfortunately, may not be available for all areas. Additionally, the manual surveys quickly go out of date within a few years of publication as additional papers are published in the field.

¹Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.* 40, 3, Article 8

Thus, a system that can generate such surveys automatically would be a useful tool. Short summaries in the form of abstracts are available for individual papers, but no such information is available for scientific topics. In this paper, we explore strategies for generating and evaluating such surveys of scientific topics automatically starting from a phrase representing a topic area. We evaluate our system on a set of topics in the field of Natural Language Processing. In earlier work, (Teufel and Moens, 2002) have examined the problem of summarizing scientific articles using rhetorical analysis of sentences. Nanba and Okumura (1999) have also discussed the problem of generating surveys of multiple papers. Mohammad et al. (2009) presented experiments on generating surveys of scientific topics starting from papers to be summarized. More recently, Hoang and Kan (2010) have presented initial results on automatically generating related work section for a target paper by taking a hierarchical topic tree as an input.

In this paper, we tackle the more challenging problem of summarizing a topic starting from a topic query. Our system takes as an input a string describing the topic area, selects the relevant papers from a corpus of papers, and then selects sentences from the citing sentences to these papers to generate a survey of the topic. A sample output of our system for the topic of “Word Sense Disambiguation” is shown in Figure 1.

2 Candidate Document Selection

Given a query representing the topic to be summarized, our first task is to find the set of relevant documents from the corpus. The simplest way to do this for a corpus of scientific publications is to do a query search using exact match or a standard TF*IDF system such Lucene, rank the documents using either citation counts or pagerank in the bibliometric citation network, and select the top n documents. However, comparing

Many corpus based methods have been proposed to deal with the sense disambiguation problem when given definition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g., supervised sense disambiguation (Leacock et al., 1998), and semi-supervised sense disambiguation (Yarowsky, 1995).

Most researchers working on word sense disambiguation (WSD) use manually sense tagged data such as SemCor (Miller et al., 1993) to train statistical classifiers, but also use the information in SemCor on the overall sense distribution for each word as a backoff model.

Yarowsky (1995) has proposed a bootstrapping method for word sense disambiguation.

Training of WSD Classifier Much research has been done on the best supervised learning approach for WSD (Florian and Yarowsky, 2002; Lee and Ng, 2002; Mihalcea and Moldovan, 2001; Yarowsky et al., 2001).

For example, the use of parallel corpora for sense tagging can help with word sense disambiguation (Brown et al., 1991; Dagan, 1991; Dagan and Itai, 1994; Ide, 2000; Resnik and Yarowsky, 1999).

Figure 1: A sample output survey of our system on the topic of “Word Sense Disambiguation” produced by paper selection using Restricted Expansion and sentence selection using Lexrank. In our evaluations, this survey achieved a pyramid score of 0.82 and Unnormalized RU score of 0.31.

Document selection algorithm	CG_5	CG_{10}	CG_{20}
Title match sorted with citation count	1.82	2.75	3.29
Title match sorted with pagerank	1.77	2.55	3.34
Citation expansion sorted with citation count	0.53	1.20	2.29
Citation expansion sorted with pagerank	0.20	0.78	1.99
TF*IDF ranked	0.14	0.14	0.56
TF*IDF sorted with citation count	0.44	2.25	3.18
TF*IDF sorted with pagerank	1.54	2.22	2.85
Restricted Expansion	2.52	3.91	6.01

Table 1: Comparison of different methods for document selection by measuring the Cumulative Gain (CG) of top 5, 10 and 20 results.

the results of these techniques with the papers covered by gold standard surveys on a few topics, we found that some important papers are missed by these simple approaches. One reason for this is that early papers in a field might use non-standard terms in the absence of a stable, accepted terminology. Some early Word Sense Disambiguation papers, for example, refer to the problem as Lexical Ambiguity Resolution. Additionally, papers might use alternative forms or abbreviations of topics in their titles and abstracts, e.g. for input query “Semantic Role Labelling”, papers such as (Dahlmeier et al., 2009) titled “Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases” and (Che and Liu, 2010) titled “Jointly Modeling WSD and SRL with Markov Logic” might be missed.

To find these papers, we add a simple heuristic called *Restricted Expansion*. In this method, we first create a base set B , by finding papers with an exact match to the query. This is a high precision set since a paper with a title that contains the exact query phrase is very likely to be relevant to the topic. We then find additional papers by expanding in the citation network around B , that is, by finding all the papers that are cited by or cite the papers in B , to create an extended set E . From this combined set ($B \cup E$), we create a new set F

by filtering out the set of papers that are not cited by or cite a minimum threshold t_{init} of papers in B . If the total number of papers is lower than f_{min} or higher than f_{max} , we iteratively increase or decrease t till $f_{min} \leq |F| \leq f_{max}$. This method allows us to increase our recall without losing precision. The values for our current experiments are: $t_{init} = 5$, $f_{min} = 150$, $f_{max} = 250$.

Authors	Year	Size
Surveys		
ACL Wiki	2012	4
Roberto Navigli	2009	68
Eneko Agirre; Philip Edmonds	2006	28
Xiaohua Zhou; Hyouil Han	2005	6
Nancy Ide; Jean Vronis	1998	41
Tutorials		
Sanda Harabagiu	2011	45
Diana McCarthy	2011	120
Philipp Koehn	2008	17
Rada Mihalcea	2005	186

Table 2: The set of surveys and tutorials collected for the topic of “Word Sense Disambiguation”. Sizes for surveys are expressed in number of pages, sizes for tutorials are expressed in number of slides.

To evaluate different methods of candidate document selection, we use Cumulative Gain (CG), where the weight for each paper is estimated by the fraction of surveys it appears in. Table 1 shows the average Cumulative Gain of top 5, 10 and 20 documents for each of eight methods we tried. Restricted Expansion outperformed every other method. Once we obtain a set of papers to be summarized, we select the top n most cited papers in the document set as the papers to be summarized, and extract the set of citing sentences S from all the papers in the document set to these n papers. S is the input for our sentence selection algorithms, described in Section 4.

Factoid	S1	S2	S3	S4	S5	T1	T2	T3	T4	Factoid Weight
definition of wsd	X	X	X	X	X	X	X	X	X	9
wordnet	X	X	X		X	X	X	X	X	8
knowledge based wsd		X	X	X	X	X		X	X	7
supervised wsd	X	X	X	X	X	X		X		7
senseval	X	X	X			X	X	X	X	7
definition of word senses	X		X	X		X		X	X	7
knowledge based wsd using machine readable dictionaries		X	X		X	X		X	X	6
unsupervised wsd		X	X	X		X	X	X		6
bootstrapping algorithms	X	X	X			X	X	X		6
supervised wsd using decision lists	X	X	X	X	X			X		6

Table 3: Top 10 factoids for the topic of “Word Sense Disambiguation” and their distribution across various data sources.

3 Evaluation Data for Survey Generation

We use the ACL Anthology Network (AAN) as the corpus for our experiments (Radev et al., 2013). We built a factoid inventory for seven topics in NLP based on manual written surveys in the following way. For each topic, we found at least 3 recent tutorials and 3 recent surveys on the topic and extracted the factoids that are covered in each of them. Table 2 shows the complete list of material collected for the topic of “Word Sense Disambiguation”. We found around 80 factoids per topic on an average. Once the factoids were extracted, each factoid was assigned a weight based on the number of documents it appears in, and any factoids with weight one were removed. Table 3 shows the top ten factoids in the topic of Word Sense Disambiguation along with their distribution across the different surveys and tutorials and final weight.

For each of the topics, we used the method described in Section 2 to create a candidate document set and extracted the candidate citing sentences to be used as the input for the content selection component. Each sentence in each topic was then annotated by a human judge against the factoid list for that topic. A sentence is allowed to have zero or more than one factoid. The human assessors were graduate students in Computer Science who have taken a basic “Natural Language Processing” course or an equivalent course. On an average, 375 citing sentences were annotated for each topic, with 2,625 sentences being annotated in total. We present all our experimental results on this large annotated corpora which is also available for download ².

4 Content Models

Once we have the set of input sentences, our system must select the sentences that should be part

²http://clair.si.umich.edu/corpora/survey_data/

of the survey. For this task, we experimented with three content models, described below.

4.1 Centroid

The centroid of a set of documents is a set of words that are statistically important to the cluster of documents. Centroid based summarization of a document set involves first creating the centroid of the documents, and then judging the salience of each document based on its similarity to the centroid of the document set. In our case, the input citing sentences represent the documents from which we extract the centroid. We use the centroid implementation from the publicly available summarization toolkit, MEAD (Radev et al., 2004).

4.2 Lexrank

LexRank (Erkan and Radev, 2004) is a network based content selection algorithm that works by first building a graph of all the documents in a cluster. The edges between corresponding nodes represent the cosine similarity between them. Once the network is built, the algorithm computes the salience of sentences in this graph based on their eigenvector centrality in the network.

4.3 C-Lexrank

C-Lexrank is another network based content selection algorithm that focuses on diversity (Qazvinian and Radev, 2008). Given a set of sentences, it first creates a network using these sentences and then runs a clustering algorithm to partition the network into smaller clusters that represent different aspects of the paper. The motivation behind the clustering is to include more diverse facts in the summary.

5 Experiments and Results

To do an evaluation of our different content selection methods, we first select the documents using our Restricted Expansion method, and then pick

Topic	Rand	Cent	LR	C-LR
Summarization	0.68	0.61	0.91	0.82
Question Answering	0.52	0.50	0.65	0.56
Word Sense Disambiguation	0.78	0.73	0.82	0.76
Named Entity Recognition	0.90	0.90	0.94	0.94
Sentiment Analysis	0.75	0.78	0.77	0.78
Semantic Role Labeling	0.78	0.79	0.88	0.94
Dependency Parsing	0.67	0.38	0.71	0.53
Average	0.72	0.68	0.81*	0.76

Table 4: Results of pyramid evaluation for each of the three methods and the random baseline on each topic.

the citing sentences to be used as the input to the summarization module as described in Section 2. Given this input, we generate 500 word summaries for each of the seven topics using the four methods: Centroid, Lexrank, C-Lexrank and a random baseline.

For each summary, we compute two evaluation metrics. The first is the Pyramid score (Nenkova and Passonneau, 2004) computed by treating the factoids as Summary Content Units (SCU’s). The Pyramid scores for each summary is shown in Table 4. The second metric is an Unnormalized Relative Utility score (Radev and Tam, 2003), computed using the factoid scores of sentences based on the method presented in (Qazvinian, 2012). We call this Unnormalized RU since we are not able to normalize the scores with human generated gold summaries. The results for Unnormalized RU are shown in Table 5. The parameter α is the RU penalty for including a redundant sentence subsumed by an earlier sentence. If the summary chooses a sentence s_i with score w_{orig} that is subsumed by an earlier summary sentence, the score is reduced as $w_{subsumed} = (\alpha * w_{orig})$. We approximate subsumption by marking a sentence s_j as being subsumed by s_i if $F_j \subset F_i$, where F_i and F_j are sets of factoids covered in each sentence.

Topic	Rand	Cent	LR	C-LR
Summarization	0.16	0.57	0.29	0.17
Question Answering	0.32	0.39	0.48	0.30
Word Sense Disambiguation	0.28	0.33	0.31	0.30
Named Entity Recognition	0.36	0.38	0.34	0.31
Sentiment Analysis	0.23	0.34	0.48	0.33
Semantic Role Labeling	0.11	0.17	0.16	0.21
Dependency Parsing	0.16	0.05	0.30	0.15
Average	0.23	0.32	0.34*	0.25

Table 5: Results of Unnormalized Relative Utility evaluation for the three methods and random baseline using $\alpha = 0.5$.

The reason for the relatively high scores for the random baseline is that our process to select the initial set of sentences eliminates many bad sen-

tences. For example, for a subset of 5 topics, the total input set contains 1508 sentences, out of which 922 of the sentences (60%) have at least one factoid. This makes it highly likely to pick good content sentences even when we are picking sentences at random.

We find that the Lexrank method outperforms other sentence selection methods on both evaluation metrics. The higher performance of Lexrank compared to Centroid is consistent with earlier published results (Erkan and Radev, 2004). The reason for the low performance of C-Lexrank as compared to Lexrank on this data set can be attributed to the fact that the input sentence set is derived from a much more diverse set of papers which can have a high diversity in lexical choice when describing the same factoid. Thus simple lexical similarity is not enough to find good clusters in this sentence set.

The lower Unnormalized RU scores compared to Pyramid scores indicate that we are selecting sentences containing highly weighted factoids, but we do not select the most informative sentences that contain a large number of factoids. This also shows that we select some redundant factoids, since Unnormalized RU contains a penalty for redundancy. This is again, explained by the fact that the simple lexical diversity based model in C-Lexrank is not able to detect the same factoids being present in two sentences. Despite these shortcomings, our system works quite well in terms of content selection for unseen topics, Figure 2 shows the top 5 sentences for the query “Conditional Random Fields”.

6 Conclusion and Future Work

In this paper, we described a pipeline for the generation of scientific surveys starting from a topic query. Our system is divided into two components. The first component finds the set of papers from the corpus relevant to the query using a simple heuristic called *Restricted Expansion*. The second component selects sentences from these papers to generate a survey of the topic. One of the main contributions of this work is a manually annotated data set for evaluating both the tasks. We collected 47 gold standard documents (surveys and tutorials) on seven topics in Natural Language Processing and extracted factoids for each topic. Each factoid is given an importance score based on the number of gold standard documents it appears in.

In recent years, conditional random fields (CRFs) (Lafferty et al., 2001) have shown success on a number of natural language processing (NLP) tasks, including shallow parsing (Sha and Pereira, 2003), named entity recognition (McCallum and Li, 2003) and information extraction from research papers (Peng and McCallum, 2004).

In natural language processing, two aspects of CRFs have been investigated sufficiently: one is to apply it to new tasks, such as named entity recognition (McCallum and Li, 2003; Li and McCallum, 2003; Settles, 2004), part-of-speech tagging (Lafferty et al., 2001), shallow parsing (Sha and Pereira, 2003), and language modeling (Roark et al., 2004); the other is to exploit new training methods for CRFs, such as improved iterative scaling (Lafferty et al., 2001), L-BFGS (McCallum, 2003) and gradient tree boosting (Dietterich et al., 2004)

NP chunks are very similar to the ones of Ramshaw and Marcus (1995).

CRFs have shown empirical successes recently in POS tagging (Lafferty et al., 2001), noun phrase segmentation (Sha and Pereira, 2003) and Chinese word segmentation (McCallum and Feng, 2003)

CRFs have been successfully applied to a number of real-world tasks, including NP chunking (Sha and Pereira, 2003), Chinese word segmentation (Peng et al., 2004), information extraction (Pinto et al., 2003; Peng and McCallum, 2004), named entity identification (McCallum and Li, 2003; Settles, 2004), and many others.

Figure 2: A sample output survey produced by our system on the topic of “Conditional Random Fields” using Restricted Expansion and Lexrank.

Additionally, we manually annotated 2,625 input sentences, about 375 sentences per topic, with the factoids extracted from the gold standard documents for each topic. Using this corpus, we presented experimental results for the performance of our document selection component and three sentence selection strategies.

Our results indicate three main directions for future work. We plan to look at better models of diversity in sentence selection, since methods based on simple lexical similarity do not seem to work well. The low factoid recall shown by low unnormalized RU scores suggests integrating the full text of papers with citation based summaries which might help us find factoids such as topic definitions that are unlikely to be present in citing sentences. A final goal would be to improve the readability and coherence of our system output.

Acknowledgments

We thank Vahed Qazvinian, Wanchen Lu, Ben King, and Shiwali Mohan for extremely useful discussions and help with the data annotation.

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained

herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

- Güneş Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.
- Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pages 427–435, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishnan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09*, pages 584–592, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, pages 926–931.
- Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '04)*.
- Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK.
- Vahed Qazvinian. 2012. *Using Collective Discourse to Generate Surveys of Scientific Paradigms*. Ph.D. thesis.
- Dragomir R. Radev and Daniel Tam. 2003. Summarization evaluation using relative utility. In *CIKM2003*, pages 508–511.
- Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam

Winkel, and Zhu Zhang. 2004. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, pages 1–26.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.