# Towards Coherent Multi-Document Summarization

**Janara Christensen, Mausam, Stephen Soderland, Oren Etzioni**
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{janara,mausam,soderlan,etzioni}@cs.washington.edu

## Abstract

This paper presents G-FLOW, a novel system for coherent extractive multi-document summarization (MDS).[1] Where previous work on MDS considered sentence selection and ordering separately, G-FLOW introduces a joint model for selection and ordering that balances coherence and salience. G-FLOW's core representation is a graph that approximates the discourse relations across sentences based on indicators including discourse cues, deverbal nouns, co-reference, and more. This graph enables G-FLOW to estimate the coherence of a candidate summary.

We evaluate G-FLOW on Mechanical Turk, and find that it generates dramatically better summaries than an extractive summarizer based on a pipeline of state-of-the-art sentence selection and reordering components, underscoring the value of our joint model.

## 1 Introduction

The goal of multi-document summarization (MDS) is to produce high quality summaries of collections of related documents. Most previous work in extractive MDS has studied the problems of sentence selection (*e.g.*, (Radev, 2004; Haghighi and Vanderwende, 2009)) and sentence ordering (*e.g.*, (Lapata, 2003; Barzilay and Lapata, 2008)) separately, but we believe that a joint model is necessary to produce coherent summaries. The intuition is simple: if the sentences in a summary are first selected—without regard to coherence—then a satisfactory ordering of the selected sentences may not exist.

[1] System and data at http://knowitall.cs.washington.edu/gflow/

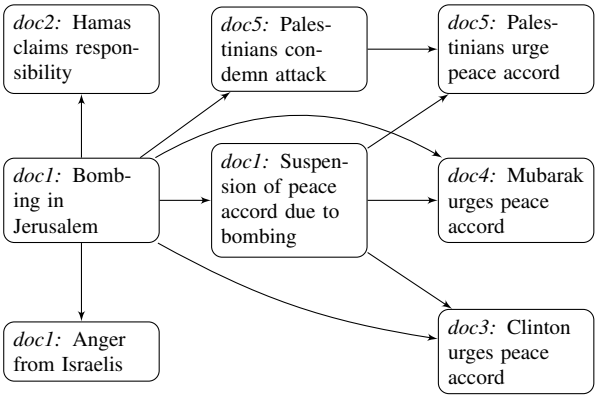


Figure 1: An example of a discourse graph covering a bombing and its aftermath, indicating the source document for each node. A coherent summary should begin with the bombing and then describe the reactions. Sentences are abbreviated for compactness.

An extractive summary is a subset of the sentences in the input documents, ordered in some way.[2] Of course, most *possible* summaries are incoherent. Now, consider a directed graph where the nodes are sentences in the collection, and each edge represents a pairwise ordering constraint necessary for a coherent summary (see Figure 1 for a sample graph). By definition, any *coherent* summary must obey the constraints in this graph.

Previous work has constructed similar graphs automatically for single document summarization and manually for MDS (see Section 2). Our system, G-FLOW extends this research in two important ways. First, it tackles automatic graph construction for MDS, which requires novel methods for identifying inter-document edges (Section 3). It uses this

[2] We focus exclusively on extractive summaries, so we drop the word "extractive" henceforth.

| State-of-the-art MDS system | G-FLOW |
|---|---|
| • The attack took place Tuesday near Cailaco in East Timor, a former Portuguese colony, according to a statement issued by the pro-independence Christian Democratic Union of East Timor.<br>• The United Nations does not recognize Indonesian claims to East Timor. | • In a decision welcomed as a landmark by Portugal, European Union leaders Saturday backed calls for a referendum to decide the fate of East Timor, the former Portuguese colony occupied by Indonesia since 1975.<br>• Indonesia invaded East Timor in 1975 and annexed it the following year. |
| • Bhichai Rattakul, deputy prime minister and president of the Bangkok Asian Games Organizing Committee, asked the Foreign Ministry to urge the Saudi government to reconsider withdrawing its 105-strong team.<br>• The games will be a success. | • Thailand won host rights for the quadrennial games in 1995, but setbacks in preparations led officials of the Olympic Council of Asia late last year to threaten to move the games to another country.<br>• Thailand showed its nearly complete facilities for the Asian Games to a tough jury Thursday - the heads of the organizing committees from the 43 nations competing in the December event. |

Table 1: Pairs of sentences produced by a pipeline of a state-of-the-art sentence extractor (Lin and Bilmes, 2011) and sentence orderer (Li et al., 2011a), and by G-FLOW.

graph to estimate coherence of a candidate summary. Second, G-FLOW introduces a novel methodology for joint sentence selection and ordering (Section 4). It casts MDS as a constraint optimization problem where salience and coherence are soft constraints, and redundancy and summary length are hard constraints. Because this optimization problem is NP-hard, G-FLOW uses local search to approximate it.

We report on a Mechanical Turk evaluation that directly compares G-FLOW to state-of-the-art MDS systems. Using DUC'04 as our test set, we compare G-FLOW against a combination of an extractive summarization system with state-of-the-art ROUGE scores (Lin and Bilmes, 2011) followed by a state-of-the-art sentence reordering scheme (Li et al., 2011a). We also compare G-FLOW to a combination of an extractive system with state-of-the-art coherence scores (Nobata and Sekine, 2004) followed by the reordering system. In both cases participants substantially preferred G-FLOW. Participants chose G-FLOW 54% of the time when compared to Lin, and chose Lin's system 22% of the time. When compared to Nobata, participants chose G-FLOW 60% of the time, and chose Nobata only 20% of the time. The remainder of the cases were judged equivalent.

A further analysis shows that G-FLOW's summaries are judged superior along several dimensions suggested in the DUC'04 evaluation (including coherence, repetitive text, and referents). A comparison against manually written, gold standard summaries, reveals that while the gold standard summaries are preferred in direct comparisons, G-FLOW has nearly equivalent scores on almost all dimensions suggested in the DUC'04 evaluation.

The paper makes the following contributions:

- We present G-FLOW, a novel MDS system that jointly solves the sentence selection and ordering problems to produce coherent summaries.
- G-FLOW automatically constructs a domain-independent graph of ordering constraints over sentences in a document collection, based on syntactic cues and redundancy across documents. This graph is the backbone for estimating the coherence of a summary.
- We perform human evaluation on blind test sets and find that G-FLOW dramatically outperforms state-of-the-art MDS systems.

## 2 Related Work

Most existing research in multi-document summarization (MDS) focuses on sentence selection for increasing coverage and does not consider coherence of the summary (Section 2.1). Although coherence has been used in ordering of summary sentences (Section 2.2), this work is limited by the quality of summary sentences given as input. In contrast, G-FLOW incorporates coherence in both selection and ordering of summary sentences.

G-FLOW can be seen as an instance of discourse-driven summarization (Section 2.3). There is prior work in this area, but primarily for summarization of single documents. There is some preliminary work on the use of manually-created discourse models in MDS. Our approach is fully automated.

### 2.1 Subset Selection in MDS

Most extractive summarization research aims to increase the coverage of concepts and entities while reducing redundancy. Approaches include the use of maximum marginal relevance (Carbonell and Goldstein, 1998), centroid-based summarization (Saggion and Gaizauskas, 2004; Radev et al., 2004), cov-

ering weighted scores of concepts (Takamura and Okumura, 2009; Qazvinian et al., 2010), formulation as minimum dominating set problem (Shen and Li, 2010), and use of submodularity in sentence selection (Lin and Bilmes, 2011). Graph centrality has also been used to estimate the salience of a sentence (Erkan and Radev, 2004). Approaches to content analysis include generative topic models (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010; Li et al., 2011b), and discriminative models (Aker et al., 2010).

These approaches do not consider coherence as one of the desiderata in sentence selection. Moreover, they do not attempt to organize the selected sentences into an intelligible summary. They are often evaluted by ROUGE (Lin, 2004), which is coherence-insensitive. In practice, these approaches often result in incoherent summaries.

### 2.2 Sentence Reordering

A parallel thread of research has investigated taking a set of summary sentences as input and reordering them to make the summary fluent. Various algorithms use some combination of topic-relatedness, chronology, precedence, succession, and entity coherence for reordering sentences (Barzilay et al., 2001; Okazaki et al., 2004; Barzilay and Lapata, 2008; Bollegala et al., 2010). Recent work has also used event-based models (Zhang et al., 2010) and context analysis (Li et al., 2011a).

The hypothesis in this research is that a pipelined combination of subset selection and reordering will produce high-quality summaries. Unfortunately, this is not true in practice, because sentences are selected primarily for coverage without regard to coherence. This methodology often leads to an inadvertent selection of a set of disconnected sentences, which cannot be put together in a coherent summary, irrespective of how the succeeding algorithm reorders them. In our evaluation, reordering had limited impact on the quality of the summaries.

### 2.3 Coherence Models and Summarization

Research on discourse analysis of documents provides a basis for modeling coherence in a document. Several theories have been developed for modeling discourse, *e.g.*, Centering Theory, Rhetorical Structure Theory (RST), Penn Discourse TreeBank (Grosz and Sidner, 1986; Mann and Thompson, 1988; Wolf and Gibson, 2005; Prasad et al., 2008). Numerous discourse-guided summarization algorithms have been developed (Marcu, 1997; Mani, 2001; Taboada and Mann, 2006; Barzilay and Elhadad, 1997; Louis et al., 2010). However, these approaches have been applied to single document summarization and not to MDS.

Discourse models have seen some application to summary generation in MDS, for example, using a detailed semantic representation of the source texts (McKeown and Radev, 1995; Radev and McKeown, 1998). A multi-document extension of RST is Cross-document Structure Theory (CST), which has been applied to MDS (Zhang et al., 2002; Jorge and Pardo, 2010). However, these systems require a stronger input, such as a manual CST-annotation of the set of documents. Our work can be seen as an instance of summarization based on lightweight CST. However, a key difference is that our proposed algorithm is completely automated and does not require any additional human annotation. Additionally, while incorporating coherence into selection, this work does not attempt to order the sentences coherently, while our approach performs joint selection and ordering.

Discourse models have also been used for evaluating summary quality (Barzilay and Lapata, 2008; Louis and Nenkova, 2009; Pitler et al., 2010). Finally, there is work on generating coherent summaries in specific domains, such as scientific articles (Saggion and Lapalme, 2002; Abu-Jbara and Radev, 2011) using domain-specific cues like citations. In contrast, our work generates summaries without any domain-specific knowledge. Other research has focused on identifying coherent threads of *documents* rather than sentences (Shahaf and Guestrin, 2010).

## 3 Discourse Graph

As described in Section 1, our goal is to identify pairwise ordering constraints over a set of input sentences. These constraints specify a multi-document discourse graph, which is used by G-FLOW to evaluate the coherence of a candidate summary.

In this graph $G$, each vertex is a sentence and an edge from $s_i$ to $s_j$ indicates that $s_j$ can be placed right after $s_i$ in a coherent summary. In other words, the two share a discourse relationship. In the fol-

lowing three sentences (from possibly different documents) there should be an edge from $s_1$ to $s_2$, but not between $s_3$ and the other sentences:

$s_1$ *Militants attacked a market in Jerusalem.*

$s_2$ *Arafat condemned the bombing.*

$s_3$ *The Wye River Accord was signed in Oct.*

Discourse theories have proposed a variety of relationships between sentences such as background and interpretation. RST has 17 such relations (Mann and Thompson, 1988) and PDTB has 16 (Prasad et al., 2008). While we seek to identify pairs of sentences that have a relationship, we do not attempt to label the edges with the exact relation.

We use textual cues from the discourse literature in combination with the redundancy inherent in related documents to generate edges. Because this methodology is noisy, the graph used by G-FLOW is an approximation, which we refer to as an approximate discourse graph (ADG). We first describe the construction of this graph, and then discuss the use of the graph for summary generation (Section 4).

### 3.1 Deverbal Noun Reference

Often, the main description of an event is mentioned in a verbal phrase and subsequent references use deverbal nouns (nominalization of verbs) (*e.g.*, 'attacked' and 'the attack'). In this example, the noun is derivationally related to the verb, but that is not always the case. For example, 'bombing' in $s_2$ above refers to 'attacked' in $s_1$.

We identify verb-noun pairs with this relationship as follows. First, we locate a set of candidate pairs from WordNet: for each verb $v$, we determine potential noun references $n$ using a path length of up to two in WordNet (moving from verb to noun is possible via WordNet's 'derivationally related' links).

This set captures verb-noun pairs such as ('to attack', 'bombing'), but also includes generic pairs such as ('to act', 'attack'). To filter such errors we score the candidate references. Our goal is to emphasize common pairs and to deemphasize pairs with common verbs or verbs that map to many nouns. To this end, we score pairs by $(c/p) * (c/q)$, where $c$ is the number of times the pair $(v, n)$ appears in adjacent sentences, $p$ is the number of times the verb appears, and $q$ is the number of times that $v$ appears with a different noun. We generate these statistics over a background corpus of 60,000 articles from the New York Times and Reuters, and filter out candidate pairs scoring below a threshold identified over a small training set.

We construct edges in the ADG between pairs of sentences containing these verb to noun mappings. To our knowledge, we are the first to use deverbal nouns for summarization.

### 3.2 Event/Entity Continuation

Our second indicator is related to lexical chains (Barzilay and Lapata, 2008). We add an edge in the ADG from a sentence $s_i$ to $s_j$ if they contain the same event or entity and the timestamp of $s_i$ is less than or equal to the timestamp of $s_j$ (timestamps generated with (Chang and Manning, 2012)).

### 3.3 Discourse Markers

We use 36 explicit discourse markers (*e.g.*, 'but', 'however', 'moreover') to identify edges between two adjacent sentences of a document (Marcu and Echihabi, 2002). This indicator lets us learn an edge from $s_4$ to $s_5$ below:

$s_4$ *Arafat condemned the bombing.*

$s_5$ ***However,*** *Netanyahu suspended peace talks.*

### 3.4 Inferred Edges

We exploit the redundancy of information in MDS documents to infer edges to related sentences. An edge $(s, s'')$ can be inferred if there is an existing edge $(s, s')$ and $s'$ and $s''$ express similar information. As an example, the edge $(s_6, s_7)$ can be inferred based on edge $(s_4, s_5)$:

$s_6$ *Arafat condemned the attack.*

$s_7$ *Netanyahu has suspended the talks.*

To infer edges we need an algorithm to identify sentences expressing similar information. To identify these pairs, we extract Open Information Extraction (Banko et al., 2007) relational tuples for each sentence, and we mark any pair of sentences with an equivalent relational tuple as redundant (see Section 4.3). The inferred edges allow us to propagate within-document discourse information to sentences from other documents.

### 3.5 Co-referent Mentions

A sentence $s_j$ will not be clearly understood in isolation and may need another sentence $s_i$ in its context, if $s_j$ has a general reference (*e.g.*, 'the presi-

dent') pointing to a specific entity or event in $s_i$ (*e.g.*, 'President Bill Clinton'). We construct edges based on coreference mentions, as predicted by Stanford's coreference system (Lee et al., 2011). We are able to identify syntactic edge $(s_8, s_9)$:

$s_8$ *Pres. Clinton expressed sympathy for Israel.*

$s_9$ ***He*** *said the attack should not derail the deal.*

### 3.6 Edge Weights

We weight each edge in the ADG by adding the number of distinct indicators used to construct that edge – if sentences $s$ and $s'$ have an edge because of a discourse marker and a deverbal reference, the edge weight $w_G(s, s')$ will be two. We also include negative edges in the ADG. $w_G(s, s')$ is negative if $s'$ contains a deverbal noun reference, a discourse marker, or a co-reference mention that is not fulfilled by $s$. For example, if $s'$ contains a discourse marker, and $s$ is neither the sentence directly preceding $s'$ and there is no inferred discourse link between $s$ and $s'$, then we will add a negative edge $w_G(s, s')$.

### 3.7 Preliminary Graph Evaluation

We evaluated the quality of the ADG used by G-FLOW, which is important not only for its use in MDS, but also because the ADG may be used for other applications like topic tracking and decomposing an event into sub-events. One author randomly chose 750 edges and labeled an edge correct if the pair of sentences did have a discourse relationship between them and incorrect otherwise. 62% of the edges accurately reflected a discourse relationship. Our ADG has on average 31 edges per sentence for a dataset in which each document cluster has on average 253 sentences. This evaluation includes only the positive edges.

## 4 Summary Generation

We denote a candidate summary $X$ to be a sequence of sentences $\langle x_1, x_2, \ldots, x_{|X|} \rangle$. G-FLOW's summarization algorithm searches through the space of ordered summaries and scores each candidate summary along the dimensions of coherence (Section 4.1), salience (Section 4.2) and redundancy (Section 4.3). G-FLOW returns the summary that maximizes a joint objective function (Section 4.4).

| weight | feature |
|---|---|
| -0.037 | position in document |
| 0.033 | from first three sentences |
| -0.035 | number of people mentions |
| 0.111 | contains money |
| 0.038 | sentence length $> 20$ |
| 0.137 | length of sentence |
| 0.109 | #sentences verbs appear in (any form) |
| 0.349 | #sentences common nouns appear in |
| 0.355 | #sentences proper nouns appear in |

Table 2: Linear regression features for salience.

### 4.1 Coherence

G-FLOW estimates coherence of a candidate summary via the ADG. We define coherence as the sum of edge weights between successive summary sentences. For disconnected sentence pairs, the edge weight is zero.

$$Coh(X) = \sum_{i=1..|X|-1} w_{G+}(x_i, x_{i+1}) + \lambda w_{G-}(x_i, x_{i+1})$$

$w_{G+}$ represents positive edges and $w_{G-}$ represents negative edge weights. $\lambda$ is a tradeoff coefficient for positive and negative weights, which is tuned using the methodology described in Section 4.4.

### 4.2 Salience

Salience is the inherent value of each sentence to the documents. We compute salience of a summary ($Sal(X)$) as the sum of the saliences of individual sentences ($\sum_i Sal(x_i)$).

To estimate salience of a sentence, G-FLOW uses a linear regression classifier trained on ROUGE scores over the DUC'03 dataset. The classifier uses surface features designed to identify sentences that cover important concepts. The complete list of features and learned weights is in Table 2. The classifier finds a sentence more salient if it mentions nouns or verbs that are present in more sentences across the documents. The highest ranked features are the last three – number of other sentences that mention a noun or a verb in the given sentence. We use the same procedure as in deverbal nouns for detecting verb mentions that appear as nouns in other sentences (Section 3.1).

### 4.3 Redundancy

We also wish to avoid redundancy. G-FLOW first processes each sentence with a state-of-the-art Open Information extractor OLLIE (Mausam et al., 2012), which converts a sentence into its component relational tuples of the form (arg1, relational phrase,

arg2).[3] For example, it finds (Militants, bombed, a marketplace) as a tuple from sentence $s_{12}$.

Two sentences will express redundant information if they both contain the same or synonymous component fact(s). Unfortunately, detecting synonymy even at relational tuple level is very hard. G-FLOW approximates this synonymy by considering two relational tuples synonymous if the relation phrases contain verbs that are synonyms of each other, have at least one synonymous argument, and are timestamped within a day of each other. Because the input documents cover related events, these relatively weak rules provide good performance. The same algorithm is used for inferring edges for the ADG (Section 3.4). This algorithm can detect that the following sentences express redundant information:

$s_{12}$ *Militants bombed a marketplace in Jerusalem.*

$s_{13}$ *He alerted Arafat after assailants attacked the busy streets of Mahane Yehuda.*

### 4.4 Objective Function

The objective function needs to balance coherence, salience and redundancy and also honor the given budget, *i.e.*, maximum summary length $B$. G-FLOW treats redundancy and budget as hard constraints and coherence and salience as soft. Coherence is necessarily soft as the graph is approximate. While previous MDS systems specifically maximized coverage, in preliminary experiments on a development set, we found that adding a coverage term did not improve G-FLOW's performance. We optimize:

$$\begin{aligned} \textbf{maximize:} \quad & F(x) \triangleq Sal(X) + \alpha Coh(X) - \beta|X| \\ s.t. \quad & \textstyle\sum_{i=1..|X|} len(x_i) < B \\ & \forall x_i, x_j \in X : \text{redundant}(x_i, x_j) = 0 \end{aligned}$$

Here $len$ refers to the sentence length. We add $|X|$ term (the number of sentences in the summary) to avoid picking many short sentences, which may increase coherence and salience scores at the cost of overall summary quality.

The parameters $\alpha$, $\beta$ and $\lambda$ (see Section 4.1) are tuned automatically using a grid search over a development set as follows. We manually generate *extractive* summaries for each document cluster in our development set (DUC'03) and choose the parameter setting that minimizes $|F(X_{\text{G-FLOW}}) - F(X^*)|$ summed over all document clusters. $F$ is the objective function, $X_{\text{G-FLOW}}$ is the summary produced by G-FLOW and $X^*$ is the manual summary.

This constraint optimization problem is NP hard, which can be shown by using a reduction of the longest path problem. For this reason, G-FLOW uses local search to reach an approximation of the optimum. G-FLOW employs stochastic hill climbing with random restarts as the base search algorithm. At each step, the search either adds a sentence, removes a sentence, replaces a sentence by another, or reorders a pair of sentences. The initial summary for random restarts is constructed as follows. We first pick the highest salience sentence with no incoming negative edges as the first sentence. The following sentences are probabilistically added one at a time based on the summary score up to that sentence. The initial summary is complete when there are no possible sentences left to fit within the budget. Intuitively, this heuristic chooses a good starting point by selecting a first sentence that does not rely on context and subsequent sentences that build a high scoring summary. As with all local search algorithms, this algorithm is highly scalable and can easily apply to large collections of related documents, but does not guarantee global optima.

## 5 Experiments

Because summaries are intended for human consumption we focused on human evaluations. We hired workers on Amazon Mechanical Turk (AMT) to evaluate the summaries. Our evaluation addresses the following questions: (1) how do G-FLOW summaries compare against the state-of-the-art in MDS (Section 5.2)? (2) what is G-FLOW's performance along important summarization dimensions such as coherence and redundancy (Section 5.3)? (3) how does G-FLOW perform on coverage as measured by ROUGE (Section 5.3.1)? (4) how much do the components of G-FLOW's objective function contribute to performance (Section 5.4)? (5) how do G-FLOW's summaries compare to human summaries?

### 5.1 Data and Systems

We evaluated the systems on the Task 2 DUC'04 multi-document summarization dataset. This dataset consists of 50 clusters of related documents, each of which contains 10 documents. Each cluster of doc-

[3] Available from http://ollie.cs.washington.edu

uments also includes four gold standard summaries used for evaluation. As in the DUC'04 competition, we allowed 665 bytes for each summary including spaces and punctuation. We used DUC'03 as our development set, which contains 30 document clusters, again with approximately 10 documents each.

We compared G-FLOW against four systems. The first is a recent MDS extractive summarizer, which we choose for its state-of-the-art ROUGE scores (Lin and Bilmes, 2011).[4] The second is a pipeline of Lin's system followed by a reimplementation of a state-of-the-art sentence reordering system (Li et al., 2011a). We refer to these systems as LIN and LIN-LI, respectively. This second baseline allows us to quantify the advantage of using coherence as a factor in both sentence extraction and ordering.

We also compare against the system that had the highest coherence ratings at DUC'04 (Nobata and Sekine, 2004), which we refer to as NOBATA. As this system did not preform sentence ordering on its output, we also compare against a pipeline of Nobata's system and the sentence reordering system. We refer to this system as NOBATA-LI.

Lastly, to evaluate how well the system performs against human generated summaries, we compare against the gold standard summaries provided by DUC.

### 5.2 Overall Summary Quality

Following (Haghighi and Vanderwende, 2009) and (Celikyilmaz and Hakkani-Tur, 2010), to compare overall summary quality, we asked AMT workers to compare two candidate system summaries. The workers first read a gold standard summary, followed by the two system summaries, and were then asked to choose the better summary from the pair. The system summaries were shown in a random order to remove any bias.

To ensure that workers provided high quality data we added two quality checks. First, we restricted to workers who have an overall approval rating of over 95% on AMT. Second, we asked the workers to briefly describe the main events of the summary. We manually filtered out work where this description was incorrect.

[4]We thank Lin and Bilmes for providing us with their code. Unfortunately, we were unable to obtain other recent MDS systems from their authors.

Six workers compared each pair of summaries. We recorded the scores for each cluster, and report three numbers: the percentages of clusters where a system is more often preferred over the other and the percentage where the two systems are tied. G-FLOW is preferred almost three times as often as LIN:

| G-FLOW | Indifferent | LIN |
|---|---|---|
| 56% | 24% | 20% |

Next, we compared G-FLOW and LIN-LI. Sentence reordering improves performance, but G-FLOW is still overwhelmingly preferred:

| G-FLOW | Indifferent | LIN-LI |
|---|---|---|
| 54% | 24% | 22% |

These results suggest that incorporating coherence in sentence extraction adds significant value to a summarization system. In these experiments, LIN and LIN-LI are preferred in some cases. We analyzed those summaries more carefully, and found that occasionally, G-FLOW will sacrifice a small amount of coverage for coherence, resulting in lower performance in those cases (see Section 5.3.1).

We also compared LIN and LIN-LI, and found that reordering does not improve performance by much.

| LIN-LI | Indifferent | LIN |
|---|---|---|
| 32% | 38% | 30% |

While the scores presented above represent comparisons between G-FLOW and a summarization system with state-of-the-art ROUGE scores, we also compared against a summarization system with state-of-the-art coherence scores – the system with the highest coherence scores from DUC'04, (Nobata and Sekine, 2004). We found that G-FLOW was again preferred:

| G-FLOW | Indifferent | NOBATA |
|---|---|---|
| 68% | 10% | 22% |

Adding in sentence ordering again improved the scores for the comparison system somewhat:

| G-FLOW | Indifferent | NOBATA-LI |
|---|---|---|
| 60% | 20% | 20% |

While these scores show a significant improvement over previous sytems, they do not convey how well G-FLOW compares to the gold standard – manually generated summaries. As a final experiment, we compared G-FLOW and a second, manually generated summary:

| G-FLOW | Indifferent | Gold |
|---|---|---|
| 14% | 18% | 68% |

While we were pleased that in 32% of the cases, Turkers either preferred G-FLOW or were indifferent, there is clearly a lot of room for improvement despite the gains reported over previous sytems.

### 5.3 Comparison along Summary Dimensions

A high quality summary needs to be good along several dimensions. We asked AMT workers to rate summaries using the quality questions enumerated in DUC'04 evaluation scheme.[5] These questions concern: (1) coherence, (2) useless, confusing, or repetitive text, (3) redundancy, (4) nouns, pronouns, and personal names that are not well-specified (5) entities rementioned in an overly explicit way, (6) ungrammatical sentences, and (7) formatting errors.

We evaluated G-FLOW LIN-LI and NOBATA-LI against the gold standard summaries, using the same AMT scheme as in the previous section. To assess automated performance with respect to the standards set by human summaries, we also evaluated a (different) gold standard summary for each document cluster, using the same Mechanical Turk scheme as in the previous section. The 50 summaries produced by each system were evaluated by four workers. The results are shown in Figure 2.

G-FLOW was rated significantly better than LIN-LI in all categories except 'Redundancy' and significant better than NOBATA-LI on 'Coherence' and 'Referents'. The ratings for 'Coherence', 'Referents', and 'OverlyExplicit' are not surprising given G-FLOW's focus on coherence. The results for 'UselessText' may also be due to G-FLOW's focus on coherence which ideally prevents it from getting off topic. Lastly, G-FLOW may perform better on 'Grammatical' and 'Formatting' because it tends to choose longer sentences than other systems, which are less likely to be sentence segmentation errors. There may also be some bleeding from one dimension to the other – if a worker likes one summary she may score it highly for many dimensions.

Finally, somewhat surprisingly, we find G-FLOW's performance to be nearly that of human summaries. G-FLOW is rated statistically significantly lower than the gold summaries on only 'Redundancy'. Given the results from the previous section, G-FLOW is likely performing worse on categories not conveyed in these scores, such as Coverage, which we examine next.

[5] http://duc.nist.gov/duc2004/quality.questions.txt

| System | R | F |
|---|---|---|
| NOBATA | 30.44 | 34.36 |
| Best system in DUC-04 | 38.28 | 37.94 |
| Takamura and Okumura (2009) | 38.50 | - |
| LIN | 39.35 | 38.90 |
| G-FLOW | 37.33 | 37.43 |
| Gold Standard Summaries | **40.03** | **40.03** |

Table 3: ROUGE-1 recall and F-measure results (%) on DUC-04. Some values are missing because not all systems reported both F-measure and recall.

#### 5.3.1 Coverage Evaluation using ROUGE

Most recent research has focused on the ROUGE evaluation, and thus implicitly on coverage of information in a summary. To estimate the coverage of G-FLOW, we compared the systems on ROUGE (Lin, 2004). We calculated ROUGE-1 scores for G-FLOW, LIN, and NOBATA.[6] As sentence ordering does not matter for ROUGE, we do not include LIN-LI or NOBATA-LI in this evaluation. Because our algorithm does not explicitly maximize coverage while LIN does, we expected G-FLOW to perform slightly worse than LIN.

The ROUGE-1 scores for G-FLOW, LIN, NOBATA and other recent MDS systems are listed in Table 3. We also include the ROUGE-1 scores for the gold summaries (compared to the other gold summaries). G-FLOW has slightly lower scores than LIN and the gold standard summaries, but much higher scores than NOBATA. G-FLOW only scores significantly lower than LIN and the gold standard summaries.

We can conclude that good summaries have both the characteristics listed in the quality dimensions, and good coverage. The gold standard summaries outperform G-FLOW on both ROUGE scores and the quality dimension scores, and therefore, outperform G-FLOW on overall comparison. However, G-FLOW is preferred to LIN-LI in addition to NOBATA-LI indicating that its quality scores outweigh its ROUGE scores in that comparison. An improvement to G-FLOW may focus on increasing

[6] ROUGE version 1.5.5 with options: -a -c 95 -b 665 -m -n 4 -w 1.2

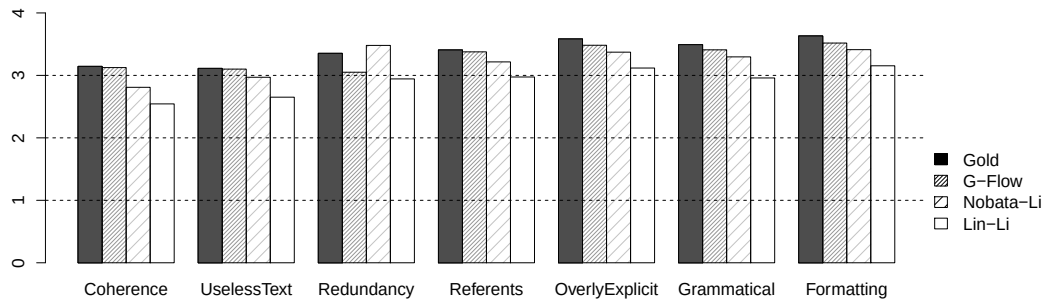


Figure 2: Ratings for the systems. 0 is the lowest possible score and 4 is the highest possible score. G-FLOW is rated significantly higher than LIN-LI on all categories, except for 'Redundancy', and significantly higher than NOBATA-LI on 'Coherence' and 'Referents'. G-FLOW is only significantly lower than the gold standard on 'Redundancy'.

coverage while retaining strengths such as coherence.

### 5.4 Ablation Experiments

In this ablation study, we evaluated the contribution of the main components of G-FLOW – coherence and salience. The details of the experiments are the same as in the experiment described in Section 5.2.

We first measured the importance of coherence in summary generation. This system G-FLOW-SAL is identical to the full system except that it does not include the coherence term in the objective function (see Section 4.4). The results show that coherence is very important to G-FLOW's performance:

| G-FLOW | Indifferent | G-FLOW-SAL |
|---|---|---|
| 54% | 26% | 20% |

Similarly, we evaluated the contribution of salience. This system G-FLOW-COH does not include the salience term in the objective function:

| G-FLOW | Indifferent | G-FLOW-COH |
|---|---|---|
| 60% | 20% | 20% |

Without salience, the system produces readable, but highly irrelevant summaries.

### 5.5 Agreement of Expert & AMT Workers

Because summary evaluation is a relatively complex task, we compared AMT workers' annotations with expert annotations from DUC'04. We randomly selected ten summaries from each of the seven DUC'04 annotators, and asked four Turk workers to annotate them on the DUC'04 quality questions. For each DUC'04 annotator, we selected all pairs of summaries where one summary was judged more than one point better than the other summary. We compared whether the workers (voting as in Section 5.2) likewise judged that summary better than the second summary. We found that the annotations agreed in 75% of cases. When we looked only at pairs more than two points different, the agreement was 80%. Thus, given the subjective nature of the task, we feel reasonably confident that the AMT annotations are informative, and that the dramatic preference of G-FLOW over the baseline systems is due to a substantial improvement in its summaries.

## 6 Conclusion

In this paper, we present G-FLOW, a multi-document summarization system aimed at generating coherent summaries. While previous MDS systems have focused primarily on salience and coverage but not coherence, G-FLOW generates an ordered summary by jointly optimizing coherence and salience. G-FLOW estimates coherence by using an approximate discourse graph, where each node is a sentence from the input documents and each edge represents a discourse relationship between two sentences. Manual evaluations demonstrate that G-FLOW generates substantially better summaries than a pipeline of state-of-the-art sentence selection and reordering components. ROUGE scores, which measure summary coverage, show that G-FLOW sacrifices a small amount of coverage for overall readability and coherence. Comparisons to gold standard summaries show that G-FLOW must improve in coverage to equal the quality of manually written summaries. We believe this research has applications to other areas of summarization such as update summarization and query based summarization, and we are interested in investigating these topics in future work.

## Acknowledgements

We thank Luke Zettlemoyer, Lucy Vanderwende, Hal Daume III, Pushpak Bhattacharyya, Chris Quirk, Erik Frey, Tony Fader, Michael Schmitz, Alan Ritter, Melissa Winstanley, and the three anonymous reviewers for helpful conversations and feedback on earlier drafts. We also thank Lin and Bilmes for providing us with the code for their system. This research was supported in part by NSF grant IIS-0803481, ONR grant N00014-11-1-0294, and DARPA contract FA8750-13-2-0019, and carried out at the University of Washington's Turing Center. This paper was also supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-10-C-7058. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. 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, AFRL, or the U.S. Government.

## References

Amjad Abu-Jbara and Dragomir R. Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of ACL 2011*, pages 500–509.

Ahmet Aker, Trevor Cohn, and Robert Gaizauskas. 2010. Multi-document summarization using A * search and discriminative training. In *Proceedings of EMNLP 2010*.

Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI 2007*, pages 68–74.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay, Noemie Elhadad, and Kathleen R McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of HLT 2001*, pages 1–7.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information Process Management*, 46(1):89–109.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR 1998*, pages 335–336.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of ACL 2010*, pages 815–824.

Angel Chang and Christopher Manning. 2012. SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of LREC 2012*.

Gunes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Barbara Grosz and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. *Proceedings of NAACL 2009*, pages 362–370.

Maria Lucia Castro Jorge and Thiago Alexandre Salgueiro Pardo. 2010. *Multi-Document Summarization: Content Selection based on CST Model (Cross-document Structure Theory)*. Ph.D. thesis, Núcleo Interinstitucional de Lingüística Computacional (NILC).

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, pages 545–552.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *CoNLL 2011 Shared Task*.

Peifeng Li, Guangxi Deng, and Qiaoming Zhu. 2011a. Using context inference to improve sentence ordering for multi-document summarization. In *Proceedings of IJCNLP 2011*, pages 1055–1061.

Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011b. Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of EMNLP 2011*, pages 1137–1146.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of ACL 2011*, pages 510–520.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Annie Louis and Ani Nenkova. 2009. Automatic summary evaluation without using human models. In *Proceedings of EMNLP 2009*, pages 306–314.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of SIGDIAL 2010*, pages 59–62.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Co, Amsterdam/Philadelphia.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL 2002*, pages 368–375.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 82–88.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of EMNLP 2012*, pages 523–534.

Kathleen McKeown and Dragomir Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of SIGIR 1995*, pages 74–82.

Chikashi Nobata and Satoshi Sekine. 2004. Crl/nyu summarization system at duc-2004. In *Proceedings of DUC 2004*.

Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Improving chronological sentence ordering by precedence relation. In *Proceedings of COLING 2004*, pages 750–756.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of ACL 2010*, pages 544–554.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of COLING 2010*, pages 895–903.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.

Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Horacio Saggion and Robert Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of DUC 2004*.

Horacio Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with sumUM. *Computational Linguistics*, 28(4):497–526.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of KDD 2010*, pages 623–632.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of Coling 2010*, pages 984–992.

Maite Taboada and William C. Mann. 2006. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–588.

Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of EACL 2009*, pages 781–789.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288.

Zhu Zhang, Sasha Blair-Goldensohn, and Dragomir R. Radev. 2002. Towards CST-enhanced summarization. In *Proceedings of AAAI 2002*, pages 439–445.

Renxian Zhang, Li Wenjie, and Lu Qin. 2010. Sentence ordering with event-enriched semantics and two-layered clustering for multi-document news summarization. In *Proceedings of COLING 2010*, pages 1489–1497.