Modelling Compression with Discourse Constraints

James Clarke and **Mirella Lapata**
School of Informatics, University of Edinburgh
2 Bucclecuch Place, Edinburgh EH8 9LW, UK
jclarke@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Sentence compression holds promise for many applications ranging from summarisation to subtitle generation and subtitle generation. The task is typically performed on isolated sentences without taking the surrounding context into account, even though most applications would operate over entire documents. In this paper we present a discourse informed model which is capable of producing document compressions that are coherent and informative. Our model is inspired by theories of local coherence and formulated within the framework of Integer Linear Programming. Experimental results show significant improvements over a state-of-the-art discourse agnostic approach.

1 Introduction

The computational treatment of sentence compression has recently attracted much attention in the literature. The task can be viewed as producing a summary of a single sentence that retains the most important information and remains grammatically correct (Jing 2000). Sentence compression is commonly expressed as a word deletion problem: given an input sentence of words $W = w_1, w_2, \ldots, w_n$, the aim is to produce a compression by removing any subset of these words (Knight and Marcu 2002).

Sentence compression can potentially benefit many applications. For example, in summarisation, a compression mechanism could improve the conciseness of the generated summaries (Jing 2000; Lin 2003). Sentence compression could be also used to automatically generate subtitles for television programs; the transcripts cannot usually be used verbatim due to the rate of speech being too high (Vandeghinste and Pan 2004). Other applications include compressing text to be displayed on small screens (Corston-Oliver 2001) such as mobile phones or PDAs, and producing audio scanning devices for the blind (Grefenstette 1998).

Most work to date has focused on a rather simple formulation of sentence compression that does not allow any rewriting operations, besides word removal. Moreover, compression is performed on isolated sentences without taking into account their surrounding context. An advantage of this simple view is that it renders sentence compression amenable to a variety of learning paradigms ranging from instantiations of the noisy-channel model (Galley and McKeown 2007; Knight and Marcu 2002; Turner and Charniak 2005) to Integer Linear Programming (Clarke and Lapata 2006a) and large-margin online learning (McDonald 2006).

In this paper we take a closer look at one of the simplifications associated with the compression task, namely that sentence reduction can be realised in isolation without making use of discourse-level information. This is clearly not true — professional abstracters often rely on contextual cues while creating summaries (Endres-Niggemeyer 1998). Furthermore, determining what information is important in a sentence is influenced by a variety of contextual factors such as the discourse topic, whether the sentence introduces new entities or events that have not been mentioned before, and the reader's background knowledge.

The simplification is also at odds with most applications of sentence compression which aim to create a shorter document rather than a single sentence. The resulting document must not only be grammat-

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1–11, Prague, June 2007.

ical but also coherent if it is to function as a replacement for the original. However, this cannot be guaranteed without knowing how the discourse progresses from sentence to sentence. To give a simple example, a contextually aware compression system could drop a word or phrase from the current sentence, simply because it is not mentioned anywhere else in the document and is therefore deemed unimportant. Or it could decide to retain it for the sake of topic continuity.

We are interested in creating a compression model that is appropriate for documents and sentences. To this end, we assess whether discourse-level information is helpful. Our analysis is informed by two popular models of discourse, Centering Theory (Grosz et al. 1995) and lexical chains (Morris and Hirst 1991). Both approaches model *local coherence* — the way adjacent sentences bind together to form a larger discourse. Our compression model is an extension of the integer programming formulation proposed by Clarke and Lapata (2006a). Their approach is conceptually simple: it consists of a scoring function coupled with a small number of syntactic and semantic constraints. Discourse-related information can be easily incorporated in the form of additional constraints. We employ our model to perform sentence compression throughout a whole document (by compressing sentences sequentially) and evaluate whether the resulting text is understandable and informative using a question-answering task. Our method yields significant improvements over a discourse agnostic state-of-the-art compression model (McDonald 2006).

2 Related Work

Sentence compression has been extensively studied across different modelling paradigms and has received both generative and discriminative formulations. Most generative approaches (Galley and McKeown 2007; Knight and Marcu 2002; Turner and Charniak 2005) are instantiations of the noisy-channel model, whereas discriminative formulations include decision-tree learning (Knight and Marcu 2002), maximum entropy (Riezler et al. 2003), support vector machines (Nguyen et al. 2004), and large-margin learning (McDonald 2006). These models are trained on a parallel corpus of long *source* sentences and their *target* compressions. Using a rich feature set derived from parse trees, the models learn either which constituents to delete or which words to place adjacently in the compression output. Relatively few approaches dispense with the parallel corpus and generate compressions in an unsupervised manner using either a scoring function (Clarke and Lapata 2006a; Hori and Furui 2004) or compression rules that are approximated from a non-parallel corpus such as the Penn Treebank (Turner and Charniak 2005).

Our work differs from previous approaches in two key respects. First, we present a compression model that is contextually aware; decisions on whether to remove or retain a word (or phrase) are informed by its discourse properties (e.g., whether it introduces a new topic, whether it is semantically related to the previous sentence). Second, we apply our compression model to entire documents rather than isolated sentences. This is more in the spirit of real-world applications where the goal is to generate a condensed and coherent text. Previous work on summarisation has also utilised discourse information (e.g., Barzilay and Elhadad 1997; Daumé III and Marcu 2002; Marcu 2000; Teufel and Moens 2002). However, its application to document compression is novel to our knowledge.

3 Discourse Representation

Obtaining an appropriate representation of discourse is the first step towards creating a compression model that exploits contextual information. In this work we focus on the role of local coherence as this is prerequisite for maintaining global coherence. Ideally, we would like our compressed document to maintain the discourse flow of the original. For this reason, we automatically annotate the source document with discourse-level information which is subsequently used to inform our compression procedure. We first describe our algorithms for obtaining discourse annotations and then present our compression model.

3.1 Centering Theory

Centering Theory (Grosz et al. 1995) is an entity-orientated theory of local coherence and salience. Although an utterance in discourse may contain several entities, it is assumed that a *single entity* is salient or "centered", thereby representing the current focus. One of the main claims underlying centering is that discourse segments in which successive utterances contain common centers are more coherent than segments where the center repeatedly changes.

Each utterance U_i in a discourse segment has a list of *forward-looking centers*, $C_f(U_i)$ and a *unique backward-looking center*, $C_b(U_i)$. $C_f(U_i)$ represents a ranking of the entities invoked by U_i according to their salience. The C_b of the current utterance U_i, is the highest-ranked element in $C_f(U_{i-1})$ that is also in U_i. The C_b thus links U_i to the previous discourse, but it does so *locally* since $C_b(U_i)$ is chosen from U_{i-1}.

Centering Algorithm So far we have presented centering without explicitly stating how the concepts "utterance", "entities" and "ranking" are instantiated. A great deal of research has been devoted into fleshing these out and many different instantiations have been developed in the literature (see Poesio et al. 2004 for details). Since our aim is to identify centers in discourse automatically, our parameter choice is driven by two considerations, robustness and ease of computation.

We therefore follow previous work (e.g., Miltsakaki and Kukich 2000) in assuming that the unit of an utterance is the sentence (i.e., a main clause with accompanying subordinate and adjunct clauses). This is in line with our compression task which also operates over sentences. We determine which entities are invoked by a sentence using two methods. First, we perform named entity identification and coreference resolution on each document using LingPipe[1], a publicly available system. Named entities and all remaining nouns are added to the C_f list. Entity matching between sentences is required to determine the C_b of a sentence. This is done using the named entity's unique identifier (as provided by LingPipe) or by the entity's surface form in the case of nouns not classified as named entities.

Entities are ranked according to their grammatical roles; subjects are ranked more highly than objects, which are in turn ranked higher than other grammatical roles (Grosz et al. 1995); ties are broken using left-to-right ordering of the grammatical roles in the sentence (Tetreault 2001). We identify grammatical roles with RASP (Briscoe and Carroll 2002). Formally, our centering algorithm is as follows (where U_i corresponds to sentence i):

[1]LingPipe can be downloaded from `http://www.alias-i.com/lingpipe/`.

1. Extract entities from U_i.
2. Create $C_f(U_i)$ by ranking the entities in U_i according to their grammatical role (subjects $>$ objects $>$ others).
3. Find the highest ranked entity in $C_f(U_{i-1})$ which occurs in $C_f(U_i)$, set the entity to be $C_b(U_i)$.

The above procedure involves several automatic steps (named entity recognition, coreference resolution, identification of grammatical roles) and will unavoidably produce some noisy annotations. So, there is no guarantee that the right C_b will be identified or that all sentences will be marked with a C_b. The latter situation also occurs in passages that contain abrupt changes in topic. In such cases, none of the entities realised in U_i will occur in $C_f(U_{i-1})$. Rather than accept that discourse information may be absent in a sentence, we turn to lexical chains as an alternative means of capturing topical content within a document.

3.2 Lexical Chains

Lexical cohesion refers to the degree of semantic relatedness observed among lexical items in a document. The term was coined by Halliday and Hasan (1976) who observed that coherent documents tend to have more related terms or phrases than incoherent ones. A number of linguistic devices can be used to signal cohesion; these range from repetition, to synonymy, hyponymy and meronymy. Lexical chains are a representation of lexical cohesion as sequences of semantically related words (Morris and Hirst 1991) and provide a useful means for describing the topic flow in discourse. For instance, a document with many different lexical chains will probably contain several topics. And main topics will tend to be represented by dense and long chains. Words participating in such chains are important for our compression task — they reveal what the document is about — and in all likelihood should not be deleted.

Lexical Chains Algorithm Barzilay and Elhadad (1997) describe a technique for text summarisation based on lexical chains. Their algorithm uses WordNet to build chains of nouns (and noun compounds). These are ranked heuristically by a score based on their length and homogeneity. A summary is then produced by extracting sentences corresponding to *strong chains*, i.e., chains whose score is two standard deviations above the average score.

Like Barzilay and Elhadad (1997), we wish to determine which lexical chains indicate the most prevalent discourse topics. Our assumption is that terms belonging to these chains are indicative of the document's main focus and should therefore be retained in the compressed output. Barzilay and Elhadad's scoring function aims to identify sentences (for inclusion in a summary) that have a high concentration of chain members. In contrast, we are interested in chains that span several sentences. We thus score chains according to the number of sentences their terms occur in. For example, the chain $\{house_3,\ home_3,\ loft_3,\ house_5\}$ (where $word_i$ denotes *word* occurring in sentence i) would be given a score of two as the terms only occur in two sentences. We assume that a chain signals a prevalent discourse topic if it occurs throughout more sentences than the average chain. The scoring algorithm is outlined more formally below:

1. Compute the lexical chains for the document.
2. $Score(Chain) = Sentences(Chain)$.
3. Discard chains if $Score(Chain) < Avg(Score)$.
4. Mark terms from the remaining chains as being the focus of the document.

We use the method of Galley and McKeown (2003) to compute lexical chains for each document.[2] This is an improved version of Barzilay and Elhadad's (1997) original algorithm.

Before compression takes place, all documents are pre-processed using the centering and lexical chain algorithms described above. In each sentence we mark the center $C_b(U_i)$ if one exists. Words (or phrases) that are present in the current sentence and function as the center in the next sentence $C_b(U_{i+1})$ are also flagged. Finally, words are marked if they are part of a prevalent chain. An example of our discourse annotation is given in Figure 1.

Bad weather dashed hopes of attempts to halt the flow$_1$ during what was seen as a lull in the lava's momentum. Experts say that even if the eruption stopped today$_2$, the pressure of lava piled up behind for six miles$_3$ would bring debris cascading down on to the town anyway. Some estimate the volcano is pouring out one million tons of debris a day$_2$, at a rate$_1$ of 15 ft$_3$ per second$_2$, from a fissure that opened in mid-December.
The Italian Army yesterday$_2$ detonated 400lb of dynamite 3,500 feet up Mount Etna's slopes.

Figure 1: Excerpt of document from our test set with discourse annotations. Centers are in double boxes; terms occurring in lexical chains are in oval boxes. Words with the same subscript are members of the same chain (e.g., *today*, *day*, *second*, *yesterday*)

4 The Compression Model

Our model is an extension of the approach put forward in Clarke and Lapata (2006a). Their work tackles sentence compression as an optimisation problem. Given a long sentence, a compression is formed by retaining the words that maximise a scoring function. The latter is essentially a language model coupled with a few constraints ensuring that the resulting output is grammatical. The language model and the constraints are encoded as linear inequalities whose solution is found using Integer Linear Programming (ILP, Vanderbei 2001; Winston and Venkataramanan 2003).

We selected this model for several reasons. First it does not require a parallel corpus and thus can be ported across domains and text genres, whilst delivering state-of-the-art results (see Clarke and Lapata 2006a for details). Second, discourse-level information can be easily incorporated by augmenting the constraint set. This is not the case for other approaches (e.g., those based on the noisy channel model) where compression is modelled by grammar rules indicating which constituents to delete in a syntactic context. Third, the ILP framework delivers a globally optimal solution by searching over the entire compression space[3] without employing heuristics or approximations during decoding.

We begin by recapping the formulation of Clarke and Lapata (2006a). Let $W = w_1, w_2, \ldots, w_n$ denote a sentence for which we wish to generate a compression. A set of binary decision variables represent whether each word w_i should be included in the

[2]The software is available from `http://www1.cs.columbia.edu/~galley/`.

[3]For a sentence of length n, there are 2^n compressions.

compression or not. Let:

$$y_i = \begin{cases} 1 & \text{if } w_i \text{ is in the compression} \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in [1 \ldots n]$$

A trigram language model forms the backbone of the compression model. The language model is formulated as an integer program with the introduction of extra decision variables indicating which *word sequences* should be retained or dropped from the compression. Let:

$$p_i = \begin{cases} 1 \text{ if } w_i \text{ starts the compression} \\ 0 \text{ otherwise} \end{cases} \forall i \in [1 \ldots n]$$

$$q_{ij} = \begin{cases} 1 \text{ if sequence } w_i, w_j \text{ ends} \\ \quad \text{the compression} & \forall i \in [1 \ldots n-1] \\ 0 \text{ otherwise} & \forall j \in [i+1 \ldots n] \end{cases}$$

$$x_{ijk} = \begin{cases} 1 \text{ if sequence } w_i, w_j, w_k & \forall i \in [1 \ldots n-2] \\ \quad \text{is in the compression} & \forall j \in [i+1 \ldots n-1] \\ 0 \text{ otherwise} & \forall k \in [j+1 \ldots n] \end{cases}$$

The objective function is expressed in Equation (1). It is the sum of all possible trigrams multiplied by the appropriate decision variable. The objective function also includes a significance score for each word multiplied by the decision variable for that word (see the last summation term in (1)). This score highlights important content words in a sentence and is defined in Section 4.1.

$$\begin{aligned} \max z = & \sum_{i=1}^{n} p_i \cdot P(w_i|\text{start}) \\ & + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} x_{ijk} \cdot P(w_k|w_i, w_j) \\ & + \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} q_{ij} \cdot P(\text{end}|w_i, w_j) \\ & + \sum_{i=1}^{n} y_i \cdot I(w_i) \end{aligned} \tag{1}$$

subject to:

$$y_i, p_i, q_{ij}, x_{ijk} = 0 \text{ or } 1 \tag{2}$$

A set of *sequential* constraints[4] are added to the problem to only allow results which combine valid trigrams.

[4] We have omitted sequential constraints due to space limitations. The full details are given in Clarke and Lapata (2006a).

4.1 Significance Score

The significance score is an attempt at capturing the gist of a sentence. It gives more weight to content words that appear in the deepest level of embedding in the syntactic tree representing the source sentence:

$$I(w_i) = \frac{l}{N} \cdot f_i \log \frac{F_a}{F_i} \tag{3}$$

The score is computed over a large corpus where w_i is a content word (i.e., a noun or verb), f_i and F_i are the frequencies of w_i in the document and corpus respectively, and F_a is the sum of all content words in the corpus. l is the number of clause constituents above w_i, and N is the deepest level of embedding.

4.2 Sentential Constraints

The model also contains a small number of sentence-level constraints. Their aim is to preserve the meaning and structure of the original sentence as much as possible. The majority of constraints revolve around modification and argument structure and are defined over parse trees or grammatical relations. For example, the following constraint template disallows the inclusion of modifiers (e.g., nouns, adjectives) without their head words:

$$\begin{gathered} y_i - y_j \geq 0 \\ \forall i, j : w_j \text{ modifies } w_i \end{gathered} \tag{4}$$

Other constraints force the presence of modifiers when the head is retained in the compression. This way, it is ensured that negation will be preserved in the compressed output:

$$\begin{gathered} y_i - y_j = 0 \\ \forall i, j : w_j \text{ modifies } w_i \wedge w_j = \text{not} \end{gathered} \tag{5}$$

Argument structure constraints make sure that the resulting compression has a canonical argument structure. For instance a constraint ensures that if a verb is present in the compression then so are its arguments:

$$\begin{gathered} y_i - y_j = 0 \\ \forall i, j : w_j \in \text{subject/object of verb } w_i \end{gathered} \tag{6}$$

Finally, Clarke and Lapata (2006a) propose one discourse constraint which forces the system to preserve personal pronouns in the compressed output:

$$\begin{gathered} y_i = 1 \\ \forall i : w_i \in \text{ personal pronouns} \end{gathered} \tag{7}$$

4.3 Discourse Constraints

In addition to the constraints described above, our model includes constraints relating to the centering and lexical chains representations discussed in Section 3. Recall that after some pre-processing, each sentence is marked with: its own center $C_b(U_i)$, the center $C_b(U_{i+1})$ of the sentence following it and words that are members of high scoring chains corresponding to the document's focus. We introduce two new types of constraints based on these additional knowledge sources.

The first constraint is the centering constraint which operates over adjacent sentences. It ensures that the C_b identified in the source sentence is retained in the target compression. If present, the entity realised as the C_b in the following sentence is also retained:

$$y_i = 1 \quad \forall i : w_i \in \{C_b(U_i), C_b(U_{i+1})\} \tag{8}$$

Consider for example the discourse in Figure 1. The constraints generated from Equation (8) will require the compression to retain *lava* in the first two sentences and *debris* in sentences two and three.

We also add a lexical chain constraint. This applies only to nouns which are members of prevalent chains:

$$y_i = 1 \quad \forall i : w_i \in \text{document focus lexical chain} \tag{9}$$

This constraint is complementary to the centering constraint; the sentences it applies to do not have to be adjacent and the entities under consideration are not restricted to a specific syntactic role (e.g., subject or object). See for instance the words *flow* and *rate* in Figure 1 which are members of the same chain (marked with subscript one). According to constraint (9) both words must be included in the compressed document.

The constraints just described ensure that the compressed document will retain the discourse flow of the original and will preserve terms indicative of important topics. We argue that these constraints will additionally benefit sentence-level compression, as words which are not signalled as discourse relevant can be dropped.

4.4 Applying the Constraints

Our compression system is given a (sentence separated) document as input. The ILP model just presented is then applied sequentially to all sentences to generate a compressed version of the original. We thus create and solve an ILP for every sentence.[5] In the formulation of Clarke and Lapata (2006a) a significance score (see Section 4.1) highlights which nouns and verbs to include in the compression. As far as nouns are concerned, our discourse constraints perform a similar task. Thus, when a sentence contains discourse annotations, we are inclined to trust them more and only calculate the significance score for verbs.

During development it was observed that applying all discourse constraints simultaneously (see Equations (7)–(9)) results in relatively long compressions. To counter this, we employ these constraints using a back-off strategy that relies on progressively less reliable information. Our back-off model works as follows: if centering information is present, we apply the appropriate constraints (Equation (8)). If no centers are present, we back-off to the lexical chain information using Equation (9), and in the absence of the latter we back-off to the pronoun constraint (Equation (7)). Finally, if discourse information is entirely absent from the sentence, we default to the significance score. Sentential constraints (see Section 4.2) are applied throughout irrespectively of discourse constraints. In our test data (see Section 5 for details), the centering constraint was used in 68.6% of the sentences. The model backed off to lexical chains for 13.7% of the test sentences, whereas the pronoun constraint was applied in 8.5%. Finally, the noun and verb significance score was used on the remaining 9.2%. An example of our system's output for the text in Figure 1 is given in Figure 2.

5 Experimental Set-up

In this section we present our experimental set-up. We briefly introduce the model used for comparison with our approach and give details regarding our compression corpus and parameter estimation. Finally, we describe our evaluation methodology.

[5] We use the publicly available *lp_solve* solver (http://www.geocities.com/lpsolve/).

Bad weather dashed hopes to halt the flow during what was seen as lull in lava's momentum. Experts say that even if eruption stopped, the pressure of lava piled would bring debris cascading. Some estimate volcano is pouring million tons of debris from fissure opened in mid-December. The Army yesterday detonated 400lb of dynamite.

Figure 2: System output on excerpt from Figure 1.

Comparison with state-of-the-art An obvious evaluation experiment would involve comparing the ILP model without any discourse constraints against the discourse informed model presented in this work. Unfortunately, the two models obtain markedly different compression rates[6] which renders the comparison of their outputs problematic. To put the comparison on an equal footing, we evaluated our approach against a state-of-the-art model that achieves a compression rate similar to ours without taking discourse-level information into account. McDonald (2006) formalises sentence compression in a discriminative large-margin learning framework as a classification task: pairs of words from the source sentence are classified as being adjacent or not in the target compression. A large number of features are defined over words, parts of speech, phrase structure trees and dependencies. These are gathered over adjacent words in the compression and the words in-between which were dropped.

It is important to note that McDonald (2006) is not a straw-man system. It achieves highly competitive performance compared with Knight and Marcu's (2002) noisy channel and decision tree models. Due to its discriminative nature, the model is able to use a large feature set and to optimise compression accuracy directly. In other words, McDonald's model has a head start against our own model which does not utilise a parallel corpus and has only a few constraints. The comparison of the two systems allows us to investigate whether discourse information is redundant when using a powerful sentence compression model.

Corpus Previous work on sentence compression has used almost exclusively the Ziff-Davis, a compression corpus derived automatically from document-abstract pairs (Knight and Marcu 2002). Unfortunately, this corpus is not suitable for our purposes since it consists of isolated sentences. We thus created a document-based compression corpus manually. Following Clarke and Lapata (2006a), we asked annotators to produce compressions for 82 stories (1,629 sentences) from the BNC and the LA Times Washington Post.[7] 48 documents (962 sentences) were used for training, 3 for development (63 sentences), and 31 for testing (604 sentences).

Parameter Estimation Our parameters for the ILP model followed closely Clarke and Lapata (2006a). We used a language model trained on 25 million tokens from the North American News corpus. The significance score was based on 25 million tokens from the same corpus. Our re-implementation of McDonald (2006) used an identical feature set, and a slightly modified loss function to encourage compression on our data set.[8]

Evaluation Previous studies evaluate how well-formed the automatically derived compressions are out of context. The target sentences are typically rated by naive subjects on two dimensions, grammaticality and importance (Knight and Marcu 2002). Automatic evaluation measures have also been proposed. Riezler et al. (2003) compare the grammatical relations found in the system output against those found in a gold standard using F-score which Clarke and Lapata (2006b) show correlates reliably with human judgements.

Following previous work, sentence-based compressions were evaluated automatically using F-score computed over grammatical relations which we obtained by RASP (Briscoe and Carroll 2002). Besides individual sentences, our goal was to evaluate the compressed document as whole. Our evaluation methodology was motivated by two questions: (1) are the documents readable? and (2) how much key information is preserved between the source document and its target compression? We assume here that the compressed document is to function as a replacement for the original. We can thus measure the extent to which the compressed version can be

[6]The discourse agnostic ILP model has a compression rate of 81.2%; when discourse constraints are include the rate drops to 65.4%.

[7]The corpus is available from `http://homepages.inf.ed.ac.uk/s0460084/data/`.

[8]McDonald's (2006) results are reported on the Ziff-Davis corpus.

What is posing a threat to the town? (lava)
What hindered attempts to stop the lava flow? (bad weather)
What did the Army do first to stop the lava flow? (detonate explosives)

Figure 3: Example questions with answer key.

used to find answers for questions which are derived from the original and represent its core content.

We therefore employed a question-answering evaluation paradigm which has been previously used for summarisation evaluation and text comprehension (Mani et al. 2002; Morris et al. 1992). The overall objective of our Q&A task is to determine how accurate each document (generated by different compression systems) is at answering questions. For this we require a methodology for constructing Q&A pairs and for scoring each document.

Two annotators were independently instructed to create Q&A pairs for the original documents in the test set. Each annotator read the document and then drafted no more than ten questions and answers related to its content. Annotators were asked to create factual-based questions which required an unambiguous answer; these were typically who/what/where/when/how style questions. Annotators then compared and revised their question-answer pairs to create a common agreed upon set. Revisions typically involved merging questions, rewording and simplifying questions, and in some cases splitting a question into multiple questions. Documents for which too few questions were created or for which questions or answers were too ambiguous were removed. This left an evaluation set of six documents with between five to eight concise questions per document. Some example questions corresponding to the document from Figure 1 are given in Figure 3; correct answers are shown in parentheses.

Compressed documents and their accompanying questions were presented to human subjects who were asked to provide answers as best they could. We elicited answers for six documents in three compression conditions: gold standard, using the ILP discourse model, and McDonald's (2006) model. Each participant was also asked to rate the readability of the compressed document on a seven point scale. A Latin Square design prevented participants from seeing two different compressions of the same document.

Model	CompR	F-Score
McDonald	60.1%	36.0%*
Discourse ILP	65.4%	39.6%
Gold Standard	70.3%	—

Table 1: Compression results: compression rate and relation-based F-score; * sig. diff. from Discourse ILP ($p < 0.05$ using the Student t test).

Model	Readability	Q&A
McDonald	2.6*	53.7%*†
Discourse ILP	3.0*	68.3%
Gold Standard	5.5†	80.7%

Table 2: Human Evaluation Results: average readability ratings and average percentage of questions answered correctly. *: sig. diff. from Gold Standard; †: sig. diff. from Discourse ILP.

The study was conducted remotely over the Internet. Participants were presented with a set of instructions that explained the Q&A task and provided examples. Subjects were first asked to read the compressed document and rate its readability. Questions were then presented one at a time and participants were allowed to consult the document for the answer. Once a participant had provided an answer they were not allowed to modify it. Thirty unpaid volunteers took part in our Q&A study. All were self reported native English speakers.

The answers provided by the participants were scored against the answer key. Answers were considered correct if they were identical to the answer key or subsumed by it. For instance, *Mount Etna* was considered a right answer to the first question from Figure 3. A compressed document receives a full score if subjects have answered all questions relating to it correctly.

6 Results

As a sanity check, we first assessed the compressions produced by our model and McDonald (2006) on a sentence-by-sentence basis without taking the documents into account. There is no hope for generating shorter documents if the compressed sentences are either too wordy or too ungrammatical. Table 1 shows the compression rates (CompR) for the two systems and evaluates the quality of their output using F-score based on grammatical relations. As can be seen, the Discourse ILP compressions are slightly longer than McDonald (65.4% vs. 60.1%) but closer to the human gold standard (70.3%). This is not surprising, the Discourse ILP model takes the entire document into account, and compression decisions will be slightly more conservative. The Discourse ILP's output is significantly better than McDonald in terms of F-score, indicating that discourse-level information is generally helpful. Both systems could use further improvement as inter-annotator agreement on this data yields an F-score of 65.8%.

Let us now consider the results of our document-based evaluation. Table 2 shows the mean readability ratings obtained for each system and the percentage of questions answered correctly. We used an Analysis of Variance (ANOVA) to examine the effect of compression type (McDonald, Discourse ILP, Gold Standard). The ANOVA revealed a reliable effect on both readability and Q&A. Post-hoc Tukey tests showed that McDonald and the Discourse ILP model do not differ significantly in terms of readability. However, they are significantly less readable than the gold standard ($\alpha < 0.05$). For the Q&A task we observe that our system is significantly better than McDonald ($\alpha < 0.05$) and not significantly worse than the gold standard.

These results indicate that the automatic systems lag behind the human gold standard in terms of readability. When reading entire documents, subjects are less tolerant of ungrammatical constructions. We also find out that despite relatively low readability, the documents are overall understandable. The discourse informed model generates more informative documents — the number of questions answered correctly increases by 15% in comparison to McDonald. This is an encouraging result suggesting that there may be advantages in developing compression models that exploit contextual information.

7 Conclusions and Future Work

In this paper we proposed a novel method for automatic sentence compression. Central in our approach is the use of discourse-level information which we argue is an important prerequisite for document (as opposed to sentence) compression. Our model uses integer programming for inferring globally optimal compressions in the presence of linguistically motivated constraints. Our discourse constraints aim to capture local coherence and are inspired by centering theory and lexical chains. We showed that our model can be successfully employed to produce compressed documents that preserve most of the original's core content.

Our approach to document compression differs from most summarisation work in that our summaries are fairly long. However, we believe this is the first step into understanding how compression can help summarisation. In the future, we will interface our compression model with sentence extraction. The discourse annotations can help guide the extraction method into selecting topically related sentences which can consequently be compressed together. The compression rate can be tailored through additional constraints which act on the output length to ensure precise word limits are obeyed.

We also plan to study the effect of global discourse structure (Daumé III and Marcu 2002) on the compression task. In general, we will assess the impact of discourse information more systematically by incorporating it into generative and discriminative modelling paradigms.

Acknowledgements We are grateful to Ryan McDonald for his help with the re-implementation of his system and our annotators Vasilis Karaiskos and Sarah Luger. Thanks to Simone Teufel, Alex Lascarides, Sebastian Riedel, and Bonnie Webber for insightful comments and suggestions. Lapata acknowledges the support of EPSRC (grant GR/T04540/01).

References

Barzilay, R. and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS), ACL-97*.

Briscoe, E. J. and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC–2002)*. Las Palmas, Gran Canaria, pages 1499–1504.

Clarke, James and Mirella Lapata. 2006a. Constraint-based sentence compression: An integer programming approach. In *Proceedings of the COLING/ACL 2006 Main Conference*

Poster Sessions. Sydney, Australia, pages 144–151.

Clarke, James and Mirella Lapata. 2006b. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pages 377–384.

Corston-Oliver, Simon. 2001. Text Compaction for Display on Very Small Screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*. Pittsburgh, PA, pages 89–98.

Daumé III, Hal and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. Philadelphia, PA, pages 449–456.

Endres-Niggemeyer, Brigitte. 1998. *Summarising Information*. Springer, Berlin.

Galley, Michel and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI–03)*. pages 1486–1488.

Galley, Michel and Kathleen McKeown. 2007. Lexicalized markov grammars for sentence compression. In *In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT–2007)*. Rochester, NY.

Grefenstette, Gregory. 1998. Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the Blind. In *Proceedings of the AAAI Symposium on Intelligent Text Summarization*. Stanford, CA, pages 111–117.

Grosz, Barbara J., Scott Weinstein, and Aravind K. Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Hori, Chiori and Sadaoki Furui. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems* E87-D(1):15–25.

Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the 6th conference on Applied Natural Language Processing (ANLP–2000)*. Seattle, WA, pages 310–315.

Knight, Kevin and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence* 139(1):91–107.

Lin, Chin-Yew. 2003. Improving summarization performance by sentence compression — a pilot study. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*. Sapporo, Japan, pages 1–8.

Mani, Inderjeet, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. SUMMAC: A text summarization evaluation. *Natural Language Engineering* 8(1):43–68.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, Cambridge, MA.

McDonald, Ryan. 2006. Discriminative sentence compression with soft syntactic constraints. In *Proceedings of the 11th EACL*. Trento, Italy.

Miltsakaki, Eleni and Karen Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL–2000)*. pages 408–415.

Morris, A., G. Kasper, and D. Adams. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research* 3(1):17–35.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1):21–48.

Nguyen, Minh Le, Akira Shimazu, Susumu Horiguchi, Tu Bao Ho, and Masaru Fukushi. 2004. Probabilistic sentence reduction using support vector machines. In *Proceedings of the 20th COLING*. Geneva, Switzerland, pages 743–749.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: a parametric theory and its instantiations. *Computational Linguistics* 30(3):309–363.

Riezler, Stefan, Tracy H. King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the HLT/NAACL*. Edmonton, Canada, pages 118–125.

Tetreault, Joel R. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics* 27(4):507–520.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics* 28(4):409–446.

Turner, Jenine and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd ACL*. Ann Arbor, MI, pages 290–297.

Vandeghinste, Vincent and Yi Pan. 2004. Sentence compression for automated subtitling: A hybrid approach. In *Proceedings of the ACL Workshop on Text Summarization*. Barcelona, Spain, pages 89–95.

Vanderbei, Robert J. 2001. *Linear Programming: Foundations and Extensions*. Kluwer Academic Publishers, Boston, 2nd edition.

Winston, Wayne L. and Munirpallam Venkataramanan. 2003. *Introduction to Mathematical Programming*. Brooks/Cole.