Title Generation with Quasi-Synchronous Grammar

Title Generation with Quasi-Synchronous Grammar

Kristian Woodsend, Yansong Feng and Mirella Lapata
School of Informatics, University of Edinburgh
Edinburgh EH8 9AB, United Kingdom
k.woodsend@ed.ac.uk, Y.Feng-4@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

The task of selecting information and rendering it appropriately appears in multiple contexts in summarization. In this paper we present a model that simultaneously optimizes selection and rendering preferences. The model operates over a phrase-based representation of the source document which we obtain by merging PCFG parse trees and dependency graphs. Selection preferences for individual phrases are learned discriminatively, while a quasi-synchronous grammar (Smith and Eisner, 2006) captures rendering preferences such as paraphrases and compressions. Based on an integer linear programming formulation, the model learns to generate summaries that satisfy both types of preferences, while ensuring that length, topic coverage and grammar constraints are met. Experiments on headline and image caption generation show that our method obtains state-of-the-art performance using essentially the same model for both tasks without any major modifications.

1 Introduction

Summarization is the process of condensing a source text into a shorter version while preserving its information content. Humans summarize on a daily basis and effortlessly, yet the automatic production of high-quality summaries remains a challenge.

Most work today focuses on *extractive summarization*, where a summary is created by identifying and subsequently concatenating the most important sentences in a document. The advantage of this approach is that it does not require a great deal of linguistic analysis to generate grammatical sentences, assuming the source document was well written. Unfortunately, extracts generated this way are often documents of low readability and text quality, and contain much redundant information. The conciseness can be improved when sentence extraction is interfaced with *sentence compression*, where words and clauses are deleted based on rules typically operating over parsed input (Jing, 2000; Daumé III and Marcu, 2002; Lin, 2003; Daumé III, 2006; Zajic et al., 2007; Martins and Smith, 2009).

An alternative *abstractive* or "bottom-up" approach involves identifying high-interest words and phrases in the source text, and combining them into new sentences guided by a language model (Banko et al., 2000; Soricut and Marcu, 2007). This approach has the potential to work well, breaking out of the single-sentence paradigm. Unfortunately, the resulting summaries are not always coherent — individual constituent phrases are often combined without any semantic constraints — or grammatical beyond the n-gram horizon imposed by the language model.

Constituent deletion and recombination are merely two of the many rewrite operations professional editors and abstractors employ when creating summaries (Jing, 2002). Additional operations include truncating sentences, aggregating them, and paraphrasing at word or syntax level. Furthermore, professionals write summaries in a task-specific style. News headlines for example are typically short (three to six words), written in the present tense and active voice, and often leave out forms of the verb *be*. There are also different ways of writing a headline either *directly* by stating what the docu-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 513–523,
MIT, Massachusetts, USA, 9-11 October 2010.

ment is about or *indirectly* by raising a question in the reader's mind, which the document answers.

The automatic generation of summaries similar to those produced by human abstractors is challenging because of the many constraints imposed by the task: the summary must be maximally informative and minimally redundant, grammatical, coherent, adhere to a pre-specified length and stylistic conventions. Importantly, these constraints are conflicting; the deletion of certain phrases may avoid redundancy but result in ungrammatical output and information loss.

In this paper we propose a model for summarization that attempts to capture and optimize these constraints jointly. We learn both how to select the most important information (the content), and how to render it appropriately (the style). Selection preferences are learned discriminatively, while a quasi-synchronous grammar (QG, Smith and Eisner 2006) captures rendering preferences such as paraphrases and compressions. The entire solution space of possible extractions and QG-generated paraphrases is searched efficiently through use of integer linear programming. The ILP framework allows us to model naturally as constraints, additional requirements such as sentence length, overall summary length, topic coverage and, importantly, grammaticality.

We argue that QG is attractive for describing rewrite operations common in summarization. Rather than assuming a strictly synchronous structure over the source and target sentences, QG identifies a "sloppy" alignment of parse trees assuming that the target tree is in some way "inspired by" the source tree. A key insight in our approach is to formulate the summarization problem at the *phrase* level: both QG rules and information extraction operate over individual phrases rather than (as is the norm) sentences. At this smaller unit level, QG rules become more widely applicable and compression falls naturally because only phrases deemed important should appear in the summary.

We evaluate the proposed model on headline generation and the related task of image caption generation. However, there is nothing inherent in our formulation that is specific to those two tasks; it is possible for the model to generate longer or shorter summaries, for a single or multiple documents. Experimental results show that our method obtains state-of-the-art performance, both in terms of grammaticality and informativeness for both tasks using the same summarization model.

2 Related work

Much effort in automatic summarization has been devoted to sentence extraction which is often formalized as a classification task (Kupiec et al., 1995). Given appropriately annotated training data, a binary classifier learns to predict for each document sentence if it is worth extracting. A few previous approaches have attempted to interface sentence compression with summarization. A straightforward way to achieve this is by adopting a two-stage architecture (e.g., Lin 2003) where the sentences are first extracted and then compressed or the other way round.

Other work implements a *joint* model where words are deleted and sentences selected from a document simultaneously (Daumé III and Marcu, 2002; Martins and Smith, 2009; Woodsend and Lapata, 2010). ILP models have also been developed for sentence rather than document compression (Clarke and Lapata, 2008). Dras (1999) discusses the application of ILP to reluctant paraphrasing, i.e., the task of choosing between paraphrases while conforming to length, readability, or style constraints. Again, the aim is to rewrite text without, however, content selection. Rewrite operations other than deletion tend to be hand-crafted and domain specific (Jing and McKeown, 2000). Notable exceptions are Cohn and Lapata (2008) and Zhao et al. (2009) who present a model that can both compress and paraphrase individual sentences without however generating document-level summaries.

Headline generation is a well-studied task within single-document summarization, due to its prominence in the DUC-03 and DUC-04 evaluation competitions.[1] Many approaches identify the most informative sentence in a given document (typically the first sentence for the news genre) and subsequently apply a form of sentence compression such that the headline meets some length requirement (Dorr

[1] Approaches to headline generation are too numerous to list in detail; see the proceedings of DUC-03 and DUC-04 for an overview.

et al., 2003). The compressed sentence may also be "padded" with important content words or phrases to ensure that the topic of the document is covered (Zajic et al., 2004). Other work generates headlines in a bottom-up fashion starting from important, individual words and phrases, that are glued together to create a fluent sentence. For example, Banko et al. (2000) draw inspiration from Machine Translation and generate headlines using statistical models for content selection and sentence realization.

Relatively little work has focused on caption generation, a task related to headline generation. The aim here is to create a short, title-like description of an image embedded in a news article. Like headlines, captions have to be short and informative. In addition, a good caption must clearly identify the subject of the picture and establish its relevance to the article. Feng and Lapata (2010a) develop extractive and abstractive caption generation models that operate over the output of a probabilistic image annotation model that preprocesses the pictures and suggests keywords to describe their content. Their best model is an extension of Banko et al.'s (2000) word-based model for headline generation to phrases.

Our own work develops an ILP-based summarization model with rewrite operations that are not limited to deletion, are defined over phrases, and encoded in quasi-synchronous grammar. The QG formalism has been previously applied to parser adaptation and projection (Smith and Eisner, 2009), paraphrase identification (Das and Smith, 2009), and question answering (Wang et al., 2007); however the use of QG in summarization is novel to our knowledge. Unlike most synchronous grammar formalisms, QG does not posit a strict isomorphism between a source sentence and its target translation; it only loosely links the syntactic structure of the two, and is therefore well suited to describing the relationship between a document and its abstract. We propose an ILP formulation which not only allows to efficiently search through the space of many QG rules but also to incorporate constraints relating to content, style, and the task at hand.

3 Modeling

There are three components to our model. Content selection is performed discriminatively; an SVM learns which information in the source document should be in the summary, and gives a real-valued *salience score* for each phrase. QG rules are used to generate compressions and paraphrases of the source sentences. An ILP model combines the output of these two components into an output summary, while optimizing content selection and surface realization preferences jointly.

3.1 Document Representation

Our model operates on documents annotated with syntactic information which we obtain by parsing every sentence twice, once with a phrase structure parser and once with a dependency parser. The output from the two representations is combined into a single data structure, by mapping the dependencies to the edges of the phrase structure tree. The procedure is described in detail in Woodsend and Lapata (2010). However, we do not merge the leaf nodes into phrases here, but keep the full tree structure, as we will apply compression to phrases through the QG. In our experiments, we obtain this combined representation from the output of the Stanford parser (Klein and Manning, 2003) but any other broadly similar parser could be used instead.

3.2 Quasi-synchronous grammar

Given an input sentence S1 or its parse tree T1, the QG constructs a monolingual grammar for parsing, or generating, the possible translation (or here, paraphrase) trees T2. A grammar node in the target tree T2 is modeled on a subset of nodes in the source tree, with a rather loose alignment between the trees.

In our approach, the process of learning the grammar is unsupervised. Each sentence of the source document is compared to each sentence in the target document — headline or caption, depending on the task. Using the combined PCFG-dependency tree representation described above, we build up a list of leaf node alignments based on lexical identity, after stemming and removing stop words. We align direct parent nodes where more than one child node aligns. A grammar rule is created if the all the nodes in the target tree can be explained using nodes from the

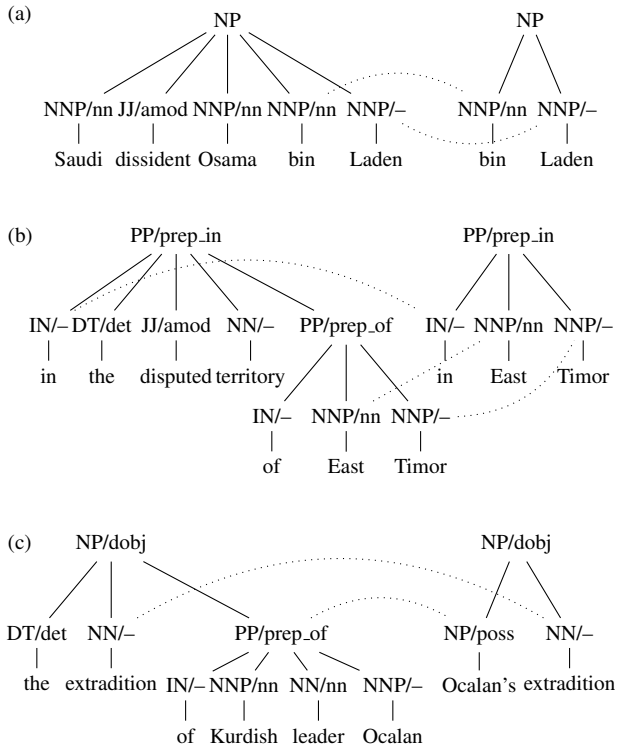

Figure 1: Examples of QG alignments between source node (left) and target node (right). (a) alignment of child nodes, involving compression through deletion; (b) rewriting involving child and grand-child nodes; (c) reordering of child nodes (with further compression through applying other QG rules on children). Nodes bear phrase and dependency labels. Dotted lines show alignments in the grammar between source and target child nodes. Examples are taken from the QG rules discovered in the DUC-03 data set of headlines.

source; this helps to improve the quality in what is inherently a noisy process. Finally, QG rules are created from aligned nodes above the leaf node level, recording the phrase and dependency label of nodes, and the alignment of child nodes.

Unlike previous work involving QG which has used dependency graphs exclusively (e.g., Wang et al. 2007; Das and Smith 2009), our approach operates over a combined PCFG-dependency representation. As a result, some configurations in Smith and Eisner (2006) are not so relevant here — instead, we found that deletions, reorderings, flattening of nodes, and the addition of text elements were important operations for the grammar.

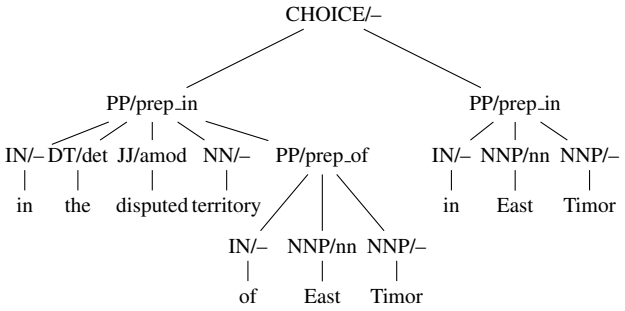

Figure 2: Alternative paraphrases are represented as a CHOICE sub-tree.

Figure 1 shows some example alignments that are captured by the QG, with the source node on the left and the target node on the right. Leaf nodes have their original text, while other nodes have a combined phrase and dependency label that they obtain in the merged representation described in Section 3.1 above (e.g., NP/dobj is a noun phrase and a direct object, NNP/nn is a proper noun and a nominal modifier, whereas NN/– is a head noun). Alignments between the children are shown by dotted lines. In Figure 1(a), some child nodes are aligned while others are not present in the target tree. This type of rule is common in our training data, and typically arises from the compression of names in noun phrases. Another frequent compression, shown in Figure 1(b), is flattening the tree structure by incorporating grand-child elements at the child level. Figure 1(c) shows a rule involving the reordering of child nodes, and where additional rules are applied recursively to achieve further compression and a transformation in the phrase constituency.

Paraphrases are created from source sentence parse trees by applying suitable rules recursively. Suitable rules have matching structure in terms of phrase and dependency label, for both the parent and child nodes. Additionally, the proposed paraphrase sub-tree must be suitable for the target tree being created (i.e., the root node of the paraphrase must match the phrase and dependency label of the corresponding node in the target tree). Where more than one paraphrase is possible, the alternatives are incorporated into the target parse tree under a *CHOICE* node, as is shown in Figure 2. Note that unlike pre-

vious QG approaches, we do not use the probability model proposed by Smith and Eisner (2006); instead the QG is used to represent rewrite operations, and we simply record a frequency count for how often each rule is encountered in the training data.

3.3 ILP model

The objective of our model is to create the most informative text possible, subject to constraints which can be tailored to the specific task. These relate to sentence length, overall summary length, the inclusion of specific topics, and grammaticality. These constraints are *global* in their scope, and cannot be adequately satisfied by optimizing each one of them individually. Our approach therefore uses an ILP formulation which will provide a globally optimal solution, and which can be efficiently solved using standard optimization tools. Specifically, the model selects phrases and paraphrases from which to form the output sentence. Here, we focus on a single sentence as this is most appropriate for title generation. However, multi-sentence output can be easily generated by setting a summary length constraint. The model operates over the merged phrase structure trees described in Section 3.1, augmented with paraphrase choice nodes such as shown in Figure 2 rather than raw text.

Let $\mathcal{S}$ be the set of sentences in a document, $\mathcal{P}$ be the set of phrases, and $\mathcal{P}_s \subset \mathcal{P}$ be the set of phrases in each sentence $s \in \mathcal{S}$. Let the sets $\mathcal{D}_i \subset \mathcal{P}, \forall i \in \mathcal{P}$ capture the phrase dependency information for each phrase i, where each set $\mathcal{D}_i$ contains the phrases that depend on the presence of i. In a similar fashion, $\mathcal{C} \subset \mathcal{P}$ is the set of choice nodes throughout the document, which represent nodes in the tree where more than one QG rule can be applied; $\mathcal{C}_i \subset \mathcal{P}, i \in \mathcal{C}$ are the sets of phrases that are direct children of each choice node, in other words they are the individual alternative paraphrases. Let l_i be the length of each phrase i, in tokens.

For caption generation, the model has as additional input a list of tags (keywords drawn from the source document) that correspond to the image, and we refer to this set of tags as $\mathcal{T}$. $\mathcal{P}_t \subset \mathcal{P}$ is the set of phrases containing the tag $t \in \mathcal{T}$. We use the probabilistic image annotation model of Feng and Lapata (2010a) to generate the list of keywords. The latter highlight the objects depicted in the image and should be in all likelihood included in the caption.

The model is cast as an integer linear program:

$$\max_{x} \quad \sum_{i \in \mathcal{P}} (f_i + \lambda g_i) x_i \tag{1a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{P}} l_i x_i \leq L_{max} \tag{1b}$$

$$\sum_{i \in \mathcal{P}} l_i x_i \geq L_{min} \tag{1c}$$

$$\sum_{i \in \mathcal{P}_t, t \in \mathcal{T}} x_i \geq T_{min} \tag{1d}$$

$$x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \tag{1e}$$

$$\sum_{j \in C_i} x_j = x_i \qquad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \tag{1f}$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \tag{1g}$$

$$\sum_{s \in \mathcal{S}} y_s \leq N_S \tag{1h}$$

$$x_i \in \{0,1\} \qquad \forall i \in \mathcal{P} \tag{1i}$$

$$y_s \in \{0,1\} \qquad \forall s \in \mathcal{S}. \tag{1j}$$

A vector of binary variables $x \in \{0,1\}^{|\mathcal{P}|}$ indicates if each phrase is to be part of the output. The vector of auxiliary binary variables $y \in \{0,1\}^{|\mathcal{S}|}$ indicates from which sentences the chosen phrases come, see Equation (1g).

Our objective function (1a) is the weighted sum of two components for each phrase: a salience score, and a measure of how frequently the QG rule was seen in the training data. Let f_i denote the salience score for phrase i, determined by the machine learning algorithm. We apply a paraphrase penalty g_i to each phrase,

$$g_i = \log\left(\frac{n_r}{N_r}\right),$$

where n_r is a count of the number of times this particular QG rule r was seen in the training data, and N_r is the number of times all suitable rules for this phrase node were seen. If no suitable rules exist, we set $g_i = 0$. The intuition here is that common paraphrases should be more trustworthy, and thus are given a smaller penalty than rare ones. Paraphrase penalties are weighted by the constant parameter λ. which controls the amount of paraphrasing we allow in the output. The objective function is the sum of the salience scores and paraphrase penalties of all the phrases chosen to form the output of a given document, subject to the constraints in Equa-

tions (1b)–(1j). The latter provide a natural way of describing the requirements the output must meet.

Constraints (1b) and (1c) ensure that the generated output stays within the acceptable length range of (L_{min}, L_{max}) tokens. Equation (1d) is a set-covering constraint, requiring that at least T_{min} words in $\mathcal{T}$ appear in the output. This is important where we want to focus on some aspect of the source document, for instance on the subject of an image.

Constraint (1e) ensures that the phrase dependencies are respected and thus enforces grammatical correctness. Phrases that depend on phrase i are contained in the set $\mathcal{D}_i$. Variable x_i is true, and therefore phrase i will be included, if any of its dependents $x_j \in \mathcal{D}_i$ are true. The phrase dependency constraints, contained in the set $\mathcal{D}_i$ and enforced by (1e), are the result of three principles based on the typed dependency information:

1. Where the QG provides alternative paraphrases, it makes sense of course to select only one. This is controlled by constraint (1f), and by placing all paraphrases in the set $\mathcal{D}_i$ for the choice node i.

2. Where there are no applicable QG rules to guide the model, in general we require all child nodes j of the current node i to be included in the summary if node i is included. As exceptions, we allow the subtree represented by node j to be deleted if the dependency label for the connecting edge $i \rightarrow j$ is of type *advcl* (adverbial clause) or some form of *conj* (conjunction).

3. In general, we force the parent node p of the current node i to be included in the output if i is, resulting in all ancestors up to the root node being included. We allow a break, and the subtree at i to be used as a stand-alone sentence, if the PCFG parser has marked i with an S (sentence) label.

Constraint (1g) tells the ILP to output a sentence if one of its constituent phrases is chosen. Finally, (1h) limits the output to a maximum of N_S sentences.

4 Experimental Set-up

As mentioned earlier we evaluated the performance of our model on two title generation tasks, namely headline and caption generation. In this section we give details on the corpora and grammars we used, model parameters and features. We also describe the baselines used for comparison with our approach, and explain how system output was evaluated.

Training We obtained phrase-based salience scores using a supervised machine learning algorithm. For the headline generation task, the full DUC-03 (Task 1) corpus was used for training; it contains 500 documents and 4 headline-style summaries per document. For the captions, training data was gathered from the CNN news website.[2] We used 200 documents and their corresponding captions. Sentences were first tokenized to separate words and punctuation, and then parsed to obtain phrases and dependencies as described in Section 3 using the Stanford parser (Klein and Manning, 2003). Document phrases were marked as positive or negative automatically. If there was a unigram overlap (excluding stop words) between the phrase and any of the original title or caption, we marked this phrase with a positive label. Non-overlapping phrases were given negative labels.

Our feature set comprised surface features such as sentence and paragraph position information, POS tags, and whether high-scoring tf.idf words were present in the phrase. Additionally, the caption training set contained features for unigram and bigram overlap with the title. We learned the feature weights with a linear SVM, using the software SVM-OOPS (Woodsend and Gondzio, 2009). This tool gave us directly the feature weights as well as support vector values, and it allowed different penalties to be applied to positive and negative misclassifications, enabling us to compensate for the unbalanced data set. The penalty hyper-parameters chosen were the ones that gave the best F-scores, using 10-fold validation.

For each of the two tasks, QG rules were extracted from the same data used to train the SVM, resulting in 2,910 distinct rules for headlines and 2,757 rules for the captions. Table 1 shows that for both tasks, the majority of rules apply to PP and NP phrases. Both tasks involve considerable compression, but the proportions of the rewrite operations involved indicate differences in style between them. Compared

[2] See `http://edition.cnn.com/`.

Label	Prop'n of set	Proportion for Label			
		Unmod	Del	Ins	Re-ord
PP	40%	5%	93%	12%	6%
NP	31%	5%	87%	14%	7%
S	20%	1%	96%	15%	7%
SBAR	6%	4%	95%	28%	6%

(a) Headlines

Label	Prop'n of set	Proportion for Label			
		Unmod	Del	Ins	Re-ord
PP	30%	17%	81%	7%	4%
NP	29%	17%	76%	11%	3%
S	27%	10%	84%	16%	6%
SBAR	10%	13%	80%	16%	3%

(b) Captions

Table 1: QG rules generated for (a) headline and (b) caption tasks (top 4 labels shown). The columns show label of root node, proportion of the full ruleset, then the proportions of rules for this label involving no modification, deletions, insertions and re-orderings.

to headlines, captions involve slightly less deletion and a higher proportion of the phrases are unmodified. The QG learning mechanism also discovers more alignments between source sentences and captions than it does for the headline task.

Title generation For the headline generation task, we evaluated our model on a testing partition from the DUC-04 corpus (75 documents, Task 1). For the caption task, we used the test set (240 documents) described in Feng and Lapata (2010a). Their corpus was downloaded from the BBC news site and contains documents, images, and their captions.[3]

We created and solved an ILP for each document. For each phrase, features were extracted and salience scores calculated from the feature weights determined through SVM training. The distance from the SVM hyperplane represents the salience score. Parameters for the ILP models for the two tasks are shown in Table 2. The λ parameter was set to 0.2 to ensure that paraphrases were included; other parameters were chosen to capture the properties seen in the majority of the training set. Note the maximum number of sentences allowed to form a headline is set to 5 as some of the headlines in the DUC dataset contained multiple sentences.

[3]Available from `http://homepages.inf.ed.ac.uk/s0677528/data.html`.

Parameter		Headlines	Captions
Min length	L_{min}	8	8
Max length	L_{max}	16	20
Min keywords	T_{min}	0	2
Max sentences	N_S	5	1
Paraphrase	λ	0.2	0.1

Table 2: ILP model parameters for the two tasks.

To solve the ILP model we used the ZIB Optimization Suite software (Achterberg, 2007; Koch, 2004). The solution was converted into a sentence by removing nodes not chosen from the tree representation, then concatenating the remaining leaf nodes in order.

Model Comparison For the headline task, we compared our model to the DUC-04 standard baseline of the first sentence, truncated at the first word boundary after 75 characters; and the output of the Topiary system (Zajic et al., 2004), which came top in almost all measures in the DUC-04 evaluation. In order to generate a headline, Topiary first compresses the lead sentence using linguistically motivated heuristics and then enhances it with topic keywords. For the captions, we compared our model against the highest-scoring document sentence according to the SVM and against the probabilistic model presented in Feng and Lapata (2010a). The latter estimates the probability of a phrase appearing in the caption given the same phrase appearing in the corresponding document and uses a language model to select among many different surface realizations. The language model is adapted with probabilities from an image annotation model (Feng and Lapata, 2010b).

Evaluation We evaluated the quality of the headlines using ROUGE (Lin and Hovy, 2003). The DUC-04 dataset provides four reference headlines per document. We report unigram overlap (ROUGE-1) and bigram overlap (ROUGE-2) as a means of assessing informativeness, and the longest common subsequence (ROUGE-L) as a means of as-

sessing fluency. Original DUC-04 ROUGE parameters were used. We also use ROUGE to evaluate the automatic captions with the original BBC captions as reference.

In addition, we evaluated the generated headlines by eliciting human judgments. Participants were presented with a news article and its corresponding headline and were asked to rate the latter along two dimensions: informativeness (does the headline capture the article's most important information?), and grammaticality (is it fluent and easy to understand?). The subjects used a seven point rating scale; an ideal system would receive high numbers for both measures. We randomly selected twelve documents from the test set and generated headlines with our model. We also included the output of Topiary and the human written DUC-04 headlines as a gold standard. We thus obtained ratings for 48 (12×4) document-highlights pairs.

We elicited judgments for the generated captions in a similar fashion. Participants were presented with a document, an associated image, and its caption, and asked to rate the latter (using a 1–7 rating scale) with respect to grammaticality and informativeness (does it describe succinctly the content of the image and document?). Again, we randomly selected 12 document-image pairs from the test set and generated captions for them using the highest scoring document sentence according to the SVM, our ILP-based model, and the output of Feng and Lapata's (2010a) system. We also included the original BBC captions as an upper bound. Both studies were conducted over the Internet using WebExp (Keller et al., 2009). 80 unpaid volunteers rated the headlines and 65 the captions, all self reported native English speakers.

5 Results

We report results on the headline generation task in Figure 3, with ROUGE-1, ROUGE-2 and ROUGE-L. In ROUGE-1 and ROUGE-L measures, the best scores are obtained by the Topiary system, slightly better than the lead sentence baseline, while for ROUGE-2 the ordering is reversed. Our model does not outperform the lead sentence or Topiary. Note that the 95% confidence level intervals reported by ROUGE are so large that no results are statistically

Lead	The chances for a new, strictly secular government in Turkey faded Wednesday.
Topiary	TURKEY YILMAZ PARTY ECEVIT chances strictly secular government faded.
ILP	Bulent Ecevit needs Turkey's two-center right parties to hammer together secular coalition.
DUC	Chance for new, secular, Turkish government fades; what will Ecevit do now?
Source	Premier-designate Bulent Ecevit needs Turkey's two-center right parties to hammer together a secular coalition, but Tansu Ciller, the ex-premier who commands 99 votes in parliament, rebuffed him Wednesday.
Lead	U.S. President Bill Clinton won South Korea's support Saturday for confronting.
Topiary	NUCLEAR U.S. President Bill Clinton won for confronting North Korea.
ILP	North Koreans have denied construction site has nuclear purpose.
DUC	U.S. warns N. Korea not to waste chance for peace over alleged nuclear site.
Source	The North Koreans have denied the underground construction site has any nuclear purpose, and it has demanded a dlrs 300 million payment for proving that.
Lead	By only one vote, the center-left prime minister of Italy, Romano Prodi.
Topiary	PRODI By only one vote center left prime minister and toppled from power.
ILP	Political system changes, Italy is condemned to political instability.
DUC	Prodi loses confidence vote; will stay as caretaker until new government.
Source	"Unless the Italian political system changes, Italy is condemned to political instability," said Sergio Romano, a former diplomat and political science professor.

Table 3: Example headline output.

F&L	The former paramedic training officer stood at the next general election.
ILP	The majority are now believing that war in Iraq was wrong.
BBC	L/Cpl Thomas Keys was shot 18 times, his inquest heard.
Source	The majority of people in this country are now believing that the war in Iraq was wrong, and I do believe we will get support.
F&L	The state government of Victoria take as those tests for cannabis.
ILP	Police in Victoria have begun randomly testing drivers for the drug ecstasy.
BBC	Police say drugs like Ecstasy can be as dangerous as alcohol for drivers.
Source	Police in the Australian state of Victoria have begun randomly testing drivers for the drug ecstasy.
F&L	The US Government Professor Holdren called for more than a year.
ILP	"We are experiencing dangerous human disruption of global climate," Professor Holdren said.
BBC	Sea levels could rise by 4m over the coming century, he warns.
Source	"We are experiencing dangerous human disruption of the global climate and we're going to experience more," Professor Holdren said.

Table 4: Example caption output.

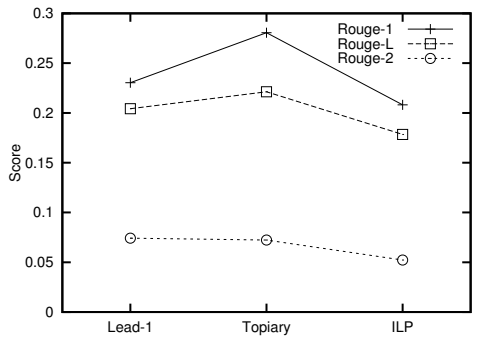

Figure 3: ROUGE-1, ROUGE-2 and ROUGE-L results on the DUC-04 headlines for our ILP model, the lead sentence baseline and Topiary.

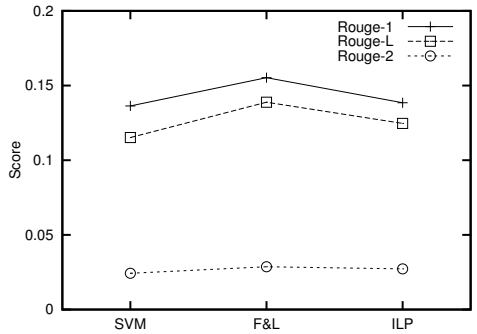

Figure 4: ROUGE-1, ROUGE-2 and ROUGE-L results on the BBC captions for our ILP model, the sentence baseline chosen by the SVM, and Feng and Lapata's (2010) model.

significant. We also investigated using an ILP model with just the QG rules or just dependency label information (see constraint (1e) in Section 3.3). Both settings gave less compressed output, and the resulting ROUGE scores were lower on all measures. The ROUGE results for the caption generation task follow a similar pattern (see Figure 4). Our model is slightly better than the best sentence baseline but performs worse than Feng and Lapata (2010a). Tables 3 and 4 show example output for the ILP model and the baselines on the headline and caption tasks respectively. In the tables, *Source* refers to the sentence chosen by the ILP, but before any paraphrasing is applied. We can see that deletion rules dominate, and a more compressive style of paraphrasing has been learned for the headline task.

The results of our human evaluation study for the DUC-04 headlines are summarized in Table 5. Means differences were compared using a Post-hoc

Model	Grammaticality	Importance
Lead-1	4.95	3.30
Topiary	3.03	3.43
ILP	5.36	4.94
Reference	5.12	5.17

Table 5: Average human ratings of DUC-04 headlines, for our ILP model, the lead sentence baseline, the output of Topiary and the human-written reference.

Model	Grammaticality	Importance
SVM	5.24	5.01
F&L	4.42	4.74
ILP	5.49	5.25
Reference	5.61	5.18

Table 6: Average human ratings of captions, for our ILP model, the sentence baseline chosen by the SVM, Feng and Lapata's (2010) model and the reference BBC caption.

Tukey test. The headlines created by our model were considered significantly more important and more grammatical than those of the Topiary system ($\alpha < 0.01$), despite the better overlap of Topiary with the reference headlines as indicated in the Rouge results above. Compared to the lead sentence of the article (the DUC-04 baseline), our model was also rated significantly higher in terms of importance ($\alpha < 0.01$) but not grammaticality.

Table 6 summarizes the results of our second judgment elicitation study. The captions generated by our model are significantly more grammatical than those of Feng and Lapata (2010a) ($\alpha < 0.01$). The SVM, ILP model and reference captions do not differ significantly in terms of grammaticality. In terms of importance, the ILP model is significantly better than the SVM ($\alpha < 0.01$) and Feng and Lapata ($\alpha < 0.01$) and comparable to the reference.

The human ratings are more favorable to our model than ROUGE for both tasks. There are two reasons for this. Firstly, the model is not biased towards selecting the lead sentence as a headline/caption and is disadvantaged in ROUGE evaluations as professional abstractors often reuse the lead or parts of it to create a title. Secondly, the model often generates an appropriate title that is lexically

distinct from the reference even though it expresses similar meaning.

6 Conclusions

In this paper we proposed a joint content selection and surface realization model for single-document summarization. The model operates over a syntax-rich representation of the source document and learns which phrases should be in the summary. Content selection preferences are coupled with a quasi-synchronous grammar whose rules encode surface realization preferences (e.g., paraphrases and compressions). Both types of preferences are optimized simultaneously in an integer linear program subject to grammaticality, length and coverage constraints. Importantly, the QG allows the model to adapt to the writing and stylistic conventions of different tasks. The results of our human studies show that our system creates grammatical and informative summaries whilst outperforming several competitive baselines.

The model itself is relatively simple and achieves good performance without any task-specific modification. One potential stumbling block may be the availability of parallel data for acquiring the QG. The Internet provides a large repository of news documents with headlines, images and captions. In some cases news articles are even accompanied with "story highlights" which could be used as training data for longer summaries.[4] For other domains obtaining such data may be more difficult. However, our experiments have shown that relatively small parallel corpora (in the range of 200–500 pairs) suffice to learn many of the writing conventions for a given task.

In the future, we plan to explore how to integrate more sophisticated QG rules in the generation process. Currently we consider deletions, reorderings and insertions. Ideally, we would also like to model arbitrary substitutions between words but also larger constituents (e.g., subclauses, sentence aggregation). Beyond summarization, we would also like to apply our model to other generation tasks, such as paraphrasing and text simplification.

[4] On-line CNN news articles are prefaced by story highlights—three or four short sentences that are written by humans and give a brief overview of the article.

Acknowledgments We are grateful to David Chiang and Noah Smith for their input on earlier versions of this work. We would also like to thank Andreas Grothey and members of ICCS at the School of Informatics for valuable discussions and comments. We acknowledge the support of EPSRC through project grants EP/F055765/1 and GR/T04540/01.

References

Achterberg, Tobias. 2007. *Constraint Integer Programming*. Ph.D. thesis, Technische Universität Berlin.

Banko, Michele, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th ACL*. Hong Kong, pages 318–325.

Clarke, James and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research* 31:399–429.

Cohn, Trevor and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd COLING*. Manchester, UK, pages 137–144.

Das, Dipanjan and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the ACL-IJCNLP*. Suntec, Singapore, pages 468–476.

Daumé III, Hal. 2006. *Practical Structured Learning Techniques for Natural Language Processing*. Ph.D. thesis, University of Southern California.

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

Dorr, Bonnie, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference*. Edmondon, Alberta, pages 1–8.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text.*. Ph.D. thesis, Macquarie University.

Feng, Yansong and Mirella Lapata. 2010a. How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 1239–1249.

Feng, Yansong and Mirella Lapata. 2010b. Topic models for image annotation and text illustration. In *Pro-*

ceedings of the NAACL HLT. Association for Computational Linguistics, Los Angeles, California, pages 831–839.

Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the 6th ANLP*. Seattle, WA, pages 310–315.

Jing, Hongyan. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics* 28(4):527–544.

Jing, Hongyan and Kathleen McKeown. 2000. Cut and paste summarization. In *Proceedings of the 1st NAACL*. Seattle, WA, pages 178–185.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. 2009. Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods* 41(1):1–12.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*. Sapporo, Japan, pages 423–430.

Koch, Thorsten. 2004. *Rapid Mathematical Prototyping*. Ph.D. thesis, Technische Universität Berlin.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of SIGIR-95*. Seattle, WA, pages 68–73.

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.

Lin, Chin-Yew and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT NAACL*. Edmonton, Canada, pages 71–78.

Martins, André and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*. Boulder, Colorado, pages 1–9.

Smith, David and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, pages 23–30.

Smith, David A. and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the EMNLP*. Suntec, Singapore, pages 822–831.

Soricut, R. and D. Marcu. 2007. Abstractive headline generation using WIDL-expressions. *Information Processing and Management* 43(6):1536–1548. Text Summarization.

Wang, Mengqiu, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the EMNLP-CoNLL*. Prague, Czech Republic, pages 22–32.

Woodsend, Kristian and Jacek Gondzio. 2009. Exploiting separability in large-scale linear support vector machine training. *Computational Optimization and Applications* Published online.

Woodsend, Kristian and Mirella Lapata. 2010. Automatic generation of story highlights. In Sandra Carberry and Stephen Clark, editors, *Proceedings of the 48th ACL*. Uppsala, Sweden, pages 565–574.

Zajic, David, Bonnie Dorr, and Richard Schwartz. 2004. BBN/UMD at DUC-2004: Topiary. In *Proceedings of the NAACL Workshop on Document Understanding*. Boston, MA, pages 112–119.

Zajic, David, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing Management Special Issue on Summarization* 43(6):1549–1570.

Zhao, Shiqi, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore, pages 834–842.