ELSEVIER

Artificial Intelligence 139 (2002) 91–107

Artificial Intelligence

www.elsevier.com/locate/artint

Summarization beyond sentence extraction: A probabilistic approach to sentence compression ☆

Kevin Knight *, Daniel Marcu

Information Sciences Institute and Department of Computer Science, University of Southern California, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA

Received 11 May 2001

Abstract

When humans produce summaries of documents, they do not simply extract sentences and concatenate them. Rather, they create new sentences that are grammatical, that cohere with one another, and that capture the most salient pieces of information in the original document. Given that large collections of text/abstract pairs are available online, it is now possible to envision algorithms that are trained to mimic this process. In this paper, we focus on sentence compression, a simpler version of this larger challenge. We aim to achieve two goals simultaneously: our compressions should be grammatical, and they should retain the most important pieces of information. These two goals can conflict. We devise both a noisy-channel and a decision-tree approach to the problem, and we evaluate results against manual compressions and a simple baseline. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Summarization; Compression; Noisy-channel model

1. Introduction

Most research in automatic summarization has focused on extraction, i.e., on identifying the most important clauses/sentences/paragraphs in texts (see [23] for a representative collection of papers). However, determining the most important textual segments is only half of what a summarization system needs to do because, in most cases, the simple

☆ This is an extended version of our paper, "Statistics-Based Summarization—Step One: Sentence Compression", which received one of the Outstanding Paper Awards at AAAI-2000, Austin, TX, USA.

* Corresponding author.
E-mail addresses: knight@isi.edu (K. Knight), marcu@isi.edu (D. Marcu).
URLs: http://www.isi.edu/~knight (K. Knight), http://www.isi.edu/~marcu (D. Marcu).

0004-3702/02/$ – see front matter
PII: S0004-3702(02)00222-9

catenation of textual segments does not yield coherent outputs. Recently, a number of researchers have started to address the problem of generating coherent summaries: McKeown et al. [26], Barzilay et al. [3], Jing and McKeown [15], Barzilay et al. [2], and Marcu and Gerber [25] in the context of multidocument summarization; and Mani et al. [22] in the context of revising single document extracts.

The approaches proposed by Witbrock and Mittal [29]; Banko et al. [1]; Berger and Mittal [5]; Jing and Hauptmann [16] are the only ones that apply a probabilistic model trained directly on ⟨*Summary, Document*⟩ pairs. However, the *Summary* here restricted to headlines, and these models have yet to scale up to generating multiple-sentence abstracts as well as well-formed, grammatical sentences.

Our goal is also to generate coherent abstracts. However, in contrast with the above work, we intend to eventually use ⟨*Abstract, Text*⟩ tuples, which are widely available, in order to automatically learn how to rewrite *Texts* as coherent *Abstracts*. In the spirit of the work in the statistical MT community, which is focused on sentence-to-sentence translations, we also decided to focus first on a simpler problem, that of *sentence compression*. We chose this problem for two reasons:

- First, the problem is complex enough to require the development of sophisticated compression models: Determining what is important in a sentence and determining how to convey the important information grammatically, using only a few words, is just a scaled down version of the text summarization problem. Yet, the problem is simple enough, since we do not have to worry yet about discourse related issues, such as coherence, anaphors, etc.
- Second, an adequate solution to this problem has an immediate impact on several applications. For example, due to time and space constraints, the generation of TV captions often requires only the most important parts of sentences to be shown on a screen [19,28]. A good sentence compression module would therefore have an impact on the task of automatic caption generation. A sentence compression module can also be used to provide audio scanning services for the blind [13], and faster access to the web from PDA devices [7]. In general, since all systems aimed at producing coherent abstracts often implement manually written sets of sentence compression rules [3,22, 26], it is likely that a good sentence compression module would impact the overall quality of these systems as well. This becomes particularly important for text genres that use long sentences.

Previous rule-based work addressing sentence compression includes Jing and Mckeown [15], Mahesh [21], Carroll et al. [9], Canning et al. [8], and Chandrasekar et al. [10].

In this paper, we present two new data-driven approaches to the sentence compression problem. Both take as input a sequence of words $W = w_1, w_2, \ldots, w_n$ (one sentence). An algorithm may drop any subset of these words. The words that remain (order unchanged) form a compression. There are 2^n compressions to choose from—some are reasonable, most are not. Our first approach develops a probabilistic noisy-channel model for sentence compression. The second approach develops a decision-based, deterministic model.

In Section 4, we evaluate both against manual compressions and a simple baseline. We present these approaches, then evaluate the algorithms and discuss how they can be extended to the problem of text compression.

2. A noisy-channel model for sentence compression

This section describes a probabilistic approach to the compression problem. In particular, we adopt the *noisy channel* framework that has been successful in a large number of other NLP applications, including speech recognition [14], machine translation [6], part-of-speech tagging [11], transliteration [17], and information retrieval [4].

In this framework, we look at a long string and imagine that (1) it was originally a short string, and then (2) someone added some additional, optional text to it. Compression is a matter of identifying the original short string. It is not critical whether or not the "original" string is real or hypothetical. For example, in statistical machine translation, we look at a French string and say, "This was originally English, but someone added 'noise' to it". The French may or may not have been translated from English originally, but by removing the noise, we can hypothesize an English source—and thereby translate the string. In the case of compression, the noise consists of optional text material that pads out the core signal. For the larger case of text summarization, it may be useful to imagine a scenario in which a news editor composes a short document, hands it to a reporter, and tells the reporter to "flesh it out" ... which results in the article we read in the newspaper. As summarizers, we may not have access to the editor's original version (which may or may not exist), but we can guess at it—which is where probabilities come in.

As in any noisy channel application, we must solve three problems:

- Source model. We must assign to every string s a probability $\mathrm{P}(s)$, which gives the chance that s is generated as an "original short string" in the above hypothetical process. For example, we may want $\mathrm{P}(s)$ to be very low if s is ungrammatical.
- Channel model. We assign to every pair of strings $\langle s, t \rangle$ a probability $\mathrm{P}(t \mid s)$, which gives the chance that when the short string s is expanded, the result is the long string t. For example, if t is the same as s except for the extra word "not", then we may want $\mathrm{P}(t \mid s)$ to be very low. The word "not" is not optional, additional material.
- Decoder. When we observe a long string t, we search for the short string s that maximizes $\mathrm{P}(s \mid t)$. This is equivalent to searching for the s that maximizes $\mathrm{P}(s) \cdot \mathrm{P}(t \mid s)$.

It is advantageous to break the problem down this way, as it decouples the somewhat independent goals of creating a short text that (1) looks grammatical, and (2) preserves important information. It is easier to build a channel model that focuses exclusively on the latter, without having to worry about the former. That is, we can specify that a certain substring may represent unimportant information, but we do not need to worry that deleting it will result in an ungrammatical structure. We leave that to the source model, which worries exclusively about well-formedness. In fact, we can make use of extensive prior work in source language modeling for speech recognition, machine translation, and

natural language generation. The same goes for actual compression ("decoding" in noisy-channel jargon)—we can re-use generic software packages to solve problems in all these application domains.

2.1. Statistical models

In the experiments we report here, we build very simple source and channel models. In a departure from the above discussion and from previous work on statistical channel models, we assign probabilities $P_{tree}(s)$ and $P_{expand\_tree}(t \mid s)$ to trees rather than strings. That is, s and t range over trees. In decoding a new string, we first parse it into a large tree t (using Collins' parser [12]), and we then hypothesize and rank various small trees.

Good source trees are ones that have both (1) a normal-looking parse structure, and (2) normal-looking word pairs. $P_{tree}(s)$ is a combination of a standard probabilistic context-free grammar (PCFG) score, which is computed over the grammar rules that yielded the tree s, and a standard word-bigram score, which is computed over the leaves of the tree. For example, the tree

```
s = S (NP John)
      (VP (VB saw)
          (NP Mary)))
```

is assigned a score based on these factors:

$$\begin{aligned} P_{tree}(s) = {} & P_{cfg}(\text{TOP} \rightarrow \text{S} \mid \text{TOP}) \cdot P_{cfg}(\text{S} \rightarrow \text{NP VP} \mid \text{S}) \cdot P_{cfg}(\text{NP} \rightarrow \text{John} \mid \text{NP}) \\ & \cdot P_{cfg}(\text{VP} \rightarrow \text{VB NP} \mid \text{VP}) \cdot P_{cfg}(\text{VP} \rightarrow \text{saw} \mid \text{VB}) \\ & \cdot P_{cfg}(\text{NP} \rightarrow \text{Mary} \mid \text{NP}) \\ & \cdot P_{bigram}(\text{John} \mid \text{EOS}) \cdot P_{bigram}(\text{saw} \mid \text{John}) \cdot P_{bigram}(\text{Mary} \mid \text{saw}) \\ & \cdot P_{bigram}(\text{EOS} \mid \text{Mary}). \end{aligned}$$

(We note that the probability assignments made by this source model do not sum to one, but it suits our purpose as it stands. Interpolating the bigram probabilities with the PCFG probabilities would be one way to straighten up the model; there are others.)

Our stochastic channel model performs minimal operations on a small tree s to create a larger tree t. For each internal node in s, we probabilistically choose an *expansion template* based on the labels of the node and its children. For example, when processing the S node in the tree above, we may wish to add a prepositional phrase as a third child. We do this with probability $P_{exp}(\text{S} \rightarrow \text{NP VP PP} \mid \text{S} \rightarrow \text{NP VP})$. Or we may choose to leave it alone, with probability $P_{exp}(\text{S} \rightarrow \text{NP VP} \mid \text{S} \rightarrow \text{NP VP})$. After we choose an expansion template, then for each new child node introduced (if any), we grow a new subtree rooted at that node—for example (PP (P in) (NP Pittsburgh)). Any particular subtree is grown with probability given by its PCFG factorization, as above (no bigrams).

This is a simple, narrow view of text expansion. This channel model only allows the insertion of new subtrees; it does not allow any sort of tree re-organization. A phrase like (NP (JJ Roman) (NN history)) cannot be expanded into (NP (NP (DT the) (NN history)) (PP (P of) (NP Rome))). To view it from the other side, imagine that we are faced with some

large tree t, and that we want to list out all small "source" trees s such that $P(t \mid s) > 0$. Then we are only allowed to delete sets of constituents in t.

2.2. Example

In this section, we show how to tell whether one potential compression is more likely than another, according to the statistical models described above. Suppose we observe the tree t in Fig. 1, which spans the string `abcde`. Consider the compression $s1$, which is shown in the same figure.

We compute the factors $P_{tree}(s1)$ and $P_{expand\_tree}(t \mid s1)$. Breaking this down further, the source PCFG that describe $P_{tree}(s1)$, are:

$$P_{cfg}(\text{TOP} \rightarrow \text{G} \mid \text{TOP}) \quad P_{cfg}(\text{G} \rightarrow \text{H A} \mid \text{G}) \quad \boxed{P_{cfg}(\text{A} \rightarrow \text{C D} \mid \text{A})}$$

$$P_{cfg}(\text{H} \rightarrow \text{a} \mid \text{H}) \qquad P_{cfg}(\text{C} \rightarrow \text{b} \mid \text{C}) \qquad P_{cfg}(\text{D} \rightarrow \text{e} \mid \text{D}).$$

The source word-bigram factors are:

$$P_{bigram}(\text{a} \mid \text{EOS}) \quad P_{bigram}(\text{b} \mid \text{a}) \quad \boxed{P_{bigram}(\text{e} \mid \text{b})} \quad P_{bigram}(\text{EOS} \mid \text{e}).$$

The channel expansion-template factors that make up part of $P_{expand\_tree}(t \mid s1)$ are:

$$P_{exp}(\text{G} \rightarrow \text{H A} \mid \text{G} \rightarrow \text{H A}) \quad \boxed{P_{exp}(\text{A} \rightarrow \text{C B D} \mid \text{A} \rightarrow \text{C D})}.$$

And finally, the "new tree growth" channel PCFG factors for the expansion are:

$$P_{cfg}(\text{B} \rightarrow \text{Q R} \mid \text{B}) \quad P_{cfg}(\text{Q} \rightarrow \text{Z} \mid \text{Q}) \quad P_{cfg}(\text{Z} \rightarrow \text{c} \mid \text{Z}) \quad P_{cfg}(\text{R} \rightarrow \text{d} \mid \text{R}).$$

A different compression will be scored with a different set of factors. For example, consider a compression of t that leaves t completely untouched. In that case, the source costs $P_{tree}(t)$ are:

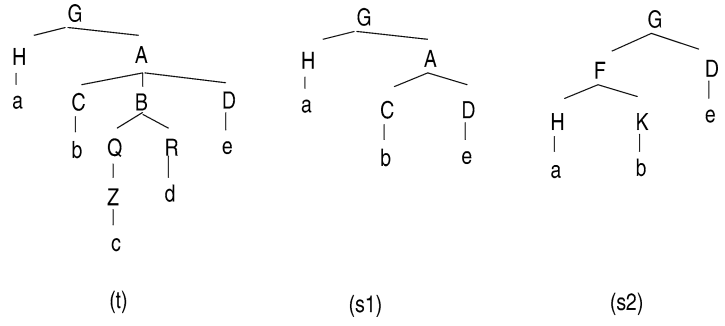

Fig. 1. Examples of parse trees.

$P_{cfg}(TOP \rightarrow G \mid TOP)$ $P_{cfg}(H \rightarrow a \mid H)$ $P_{bigram}(a \mid EOS)$
$P_{cfg}(G \rightarrow H\ A \mid G)$ $P_{cfg}(C \rightarrow b \mid C)$ $P_{bigram}(b \mid a)$

$\boxed{P_{cfg}(A \rightarrow C\ B\ D \mid A)}$ $P_{cfg}(Z \rightarrow c \mid Z)$ $\boxed{P_{bigram}(c \mid b)}$

$P_{cfg}(B \rightarrow Q\ R \mid B)$ $P_{cfg}(R \rightarrow d \mid R)$ $\boxed{P_{bigram}(d \mid c)}$

$P_{cfg}(Q \rightarrow Z \mid Q)$ $P_{cfg}(D \rightarrow e \mid D)$ $\boxed{P_{bigram}(e \mid d)}$

$P_{bigram}(EOS \mid e).$

The channel costs $P_{expand\_tree}(t \mid t)$ are:

$P_{exp}(G \rightarrow H\ A \mid G \rightarrow H\ A)$ $\boxed{P_{exp}(A \rightarrow C\ B\ D \mid A \rightarrow C\ B\ D)}$

$\boxed{P_{exp}(B \rightarrow Q\ R \mid B \rightarrow Q\ R)}$ $\boxed{P_{exp}(Q \rightarrow Z \mid Q \rightarrow Z)}$

(in other words, leave every node unchanged).

Now we can simply compare

$$P_{compress\_tree}(s1 \mid t) = P_{tree}(s1) \cdot P_{expand\_tree}(t \mid s1))/P_{tree}(t)$$

versus

$$P_{compress\_tree}(t \mid t) = P_{tree}(t) \cdot P_{expand\_tree}(t \mid t))/P_{tree}(t)$$

and select the more likely one. Note that the denominator $P_{tree}(t)$ and all the new-tree-growth PCFG factors cancel out. The quantities that differ between the two proposed compressions are boxed above. Therefore, $s1$ will be preferred over t if and only if:

$$\begin{aligned} &P_{cfg}(A \rightarrow C\ D \mid A) \cdot P_{exp}(A \rightarrow C\ B\ D \mid A \rightarrow C\ D) \cdot P_{bigram}(e \mid b) \\ &\quad > P_{cfg}(A \rightarrow C\ B\ D \mid A) \cdot P_{exp}(A \rightarrow C\ B\ D \mid A \rightarrow C\ B\ D) \\ &\qquad \cdot P_{exp}(B \rightarrow Q\ R \mid B \rightarrow Q\ R) \cdot P_{exp}(Q \rightarrow Z \mid Q \rightarrow Z) \\ &\qquad \cdot P_{bigram}(c \mid b) \cdot P_{bigram}(d \mid c) \cdot P_{bigram}(e \mid d). \end{aligned}$$

2.3. Training corpus

In order to train our system, we used the Ziff–Davis corpus, a collection of newspaper articles announcing computer products. Many of the articles in the corpus are paired with human written abstracts. We automatically extracted from the corpus a set of 1067 sentence pairs. Each pair consisted of a sentence $t = t_1, t_2, \ldots, t_n$ that occurred in the article

The *documentation is* typical of Epson quality: *excellent*.
Documentation is excellent.

All of our *design goals were achieved* and the delivered performance matches the speed of the underlying device.
All design goals were achieved.

Reach's E-mail product, *MailMan*, is a message-management system designed initially for VINES LANs that *will eventually be operating system-independent*.
MailMan will eventually be operating system-independent.

Although the modules themselves may be physically and/or electrically incompatible, the *cable-specific jacks* on them *provide industry-standard connections.*
Cable-specific jacks provide industry-standard connections.

Beyond that basic level, *the operations of the three products vary widely.*
The operations of the three products vary widely.

Ingres/Star prices start at $2,100.
Ingres/Star prices start at $2,100.

Fig. 2. Examples from our parallel corpus.

and a possibly compressed version of it $s = s_1, s_2, \ldots, s_m$, which occurred in the human written abstract. Fig. 2 shows a few sentence pairs extracted from the corpus, selected to demonstrate various types of compression by dropping words.

The possibly compressed sentence s uses the same words as the long sentence t; also, the words in the two sentences occurred in the same order. Fig. 2 shows a few sentence pairs extracted from the corpus, where the common words between the long and compressed version of the sentences are displayed in italics.

We decided to use a corpus of examples such as those shown in Fig. 2 because it is consistent with two desiderata specific to summarization work: (i) the human-written Abstract sentences are grammatical; (ii) the Abstract sentences represent in a compressed form the salient points of the original newspaper Sentences. We decided to keep in the corpus uncompressed sentences as well, since we want to learn not only *how* to compress a sentence, but also *when* to do it.

While the Ziff–Davis corpus is domain-specific, results appear to generalize, as what the model learns is mainly at the syntactic level.

2.4. Learning model parameters

We collect expansion-template probabilities from our parallel corpus. We first parse both sides of the parallel corpus, and then we identify corresponding syntactic nodes. For example, the parse tree for one sentence may begin

```
(S (NP ...)
   (VP ...)
   (PP ...))
```

while the parse tree for its compressed version may begin

(S (NP ...)
(VP ...)).

If these two S nodes are deemed to correspond, then we chalk up one joint event (S → NP VP, S → NP VP PP); afterwards we normalize so that P_{exp}(S → NP VP PP | S → NP VP) competes with other ways to expanding an S → NP VP node. Here are sample parameter values from actual training:

P_{exp}(NP → DT NN | NP → DT NN) = 0.8678
P_{exp}(NP → DT JJ NN | NP → DT NN) = 0.0287
P_{exp}(NP → DT NNP NN | NP → DT NN) = 0.0230
P_{exp}(NP → DT JJS NN | NP → DT NN) = 0.0115
P_{exp}(NP → DT NNP CD NN | NP → DT NN) = 0.0057
etc.

Not all nodes have corresponding partners; some non-correspondences are due to incorrect parses, while others are due to legitimate reformulations that are beyond the scope of our simple channel model. We use standard methods to estimate source PCFG and word-bigram probabilities (from the Penn Treebank and unannotated Wall Street Journal, respectively).

2.5. Decoding

There are vast numbers of potential compressions of a large tree t, but we can pack them all efficiently into a shared-forest structure. For each node of t that has n children, we

- generate $2^n - 1$ new nodes, one for each non-empty subset of the children, and
- pack those nodes so that they are referred to as a whole.

For example, consider the large tree t in Fig. 1. All compressions can be represented with the following rules, which encode the forest shown in Fig. 3.

G → H A	B → R	A → B C	H → a
G → H	Q → Z	A → C	C → b
G → A	A → C B D	A → B	Z → c
B → Q R	A → C B	A → D	R → d
B → Q	A → C D		D → e.

We can also assign a source PCFG and expansion-template probability to each node in the forest. For example, to the B → Q node, we can assign the expansion probability

P_{exp}(B → Q | B) · P_{exp}(B → Q R | B → Q).

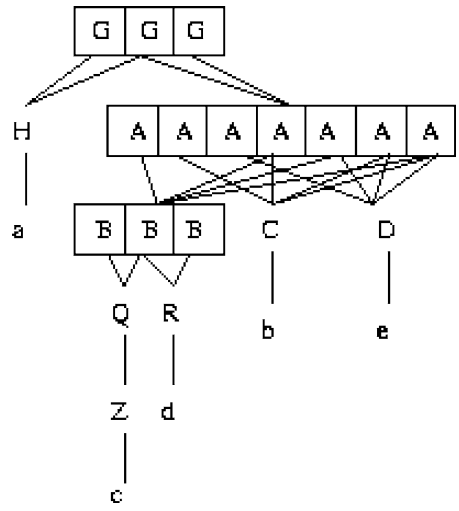

Fig. 3. A compact representation of all compressions of tree t in Fig. 1.

We smooth the expansion probabilities to 10^{-6}, but we do not smooth the source PCFG probabilities for compressed nodes. We only consider compressing a node in ways that are locally grammatical according to the Penn Treebank—e.g., if the rule A → C B has never been observed, then it will not appear in the forest.

At this point, we want to extract a set of high-scoring trees from the forest, taking into account both expansion-template probabilities and word-bigram probabilities. Fortunately, we have such a generic extractor on hand [18]. This extractor was designed for a hybrid symbolic-statistical natural language generation system called Nitrogen. In that application, a rule-based component converts an abstract semantic representation into a vast number of potential English renderings. These renderings are packed into a forest, from which the most promising sentences are extracted using statistical scoring. At present, this scoring is based on word-ngrams, but current research aims at extending the scores to account for subcategorization and long-distance syntactic relationships, as between a verb and its direct object. A more sophisticated extractor will also help determine which prepositional phrases may be safely deleted.

For our purposes, the extractor selects the trees with the best combination of word-bigram and expansion-template scores. It returns a list of such trees, one for each possible compression length. For example, for the sentence *Beyond that basic level, the operations of the three products vary*, we obtain the following "best" compressions, with negative log-probabilities shown in parentheses (smaller = more likely):

Beyond that basic level, the operations of the three products vary widely (1514588)
Beyond that level, the operations of the three products vary widely (1430374)
Beyond that basic level, the operations of the three products vary (1333437)
Beyond that level, the operations of the three products vary (1249223)
Beyond that basic level, the operations of the products vary (1181377)
The operations of the three products vary widely (939912)

The operations of the products vary widely (872066)
The operations of the products vary (748761)
The operations of products vary (690915)
Operations of products vary (809158)
The operations vary (522402)
Operations vary (662642)

2.6. Length selection

It is useful to have multiple answers to choose from, as one user may seek a 20% compression, while another seeks a 60% compression. Or, if the compression is going to be subsequently translated into another language, we may want to multiply in translation probabilities before deciding on a particular length.

However, for purposes of evaluation, we want our system to be able to select a single compression. If we rely on the log-probabilities as shown above, we will almost always choose the shortest compression. While the goal of summarization is to produce compact text, more compact is not necessarily better, as the loss of information may be too great. (Note above, however, how the three-word compression scores better than the two-word compression, as the models are not entirely happy removing the article "the".) To create a more fair competition, we divide the log-probability by the length of the compression, rewarding longer strings. This is commonly done in speech recognition (??).

If we plot this normalized score against compression length, we usually observe a (bumpy) U-shaped curve, as illustrated in Fig. 4. In a typical more difficult case, a 25-word sentence may be optimally compressed by a 17-word version. Of course, if a user requires a shorter compression than that, she may select another region of the curve and look for a local minimum.

3. A decision-based model for sentence compression

In this section, we describe a decision-based, history model of sentence compression. Decision-based models have been successful in parsing and interpretation applications Magerman [20]; Zelle and Mooney [30], and we explore their use in summarization here. As in the noisy-channel approach, we again assume that we are given as input a parse tree t. Our goal is to "rewrite" t into a smaller tree s, which corresponds to a compressed version of the original sentence subsumed by t. Suppose we observe in our corpus the trees t and $s2$ in Fig. 1. In this model, we ask ourselves how we may go about rewriting t into $s2$. One possible solution is to decompose the rewriting operation into a sequence of shift-reduce-drop actions that are specific to an extended shift-reduce parsing paradigm.

In the model we propose, the rewriting process starts with an empty Stack and an Input List that contains the sequence of words subsumed by the large tree t. Each word in the input list is labeled with the name of all syntactic constituents in t that start with it (see Fig. 5). At each step, the rewriting module applies an operation that is aimed at reconstructing the smaller tree $s2$. In the context of our sentence-compression module, we need four types of operations:

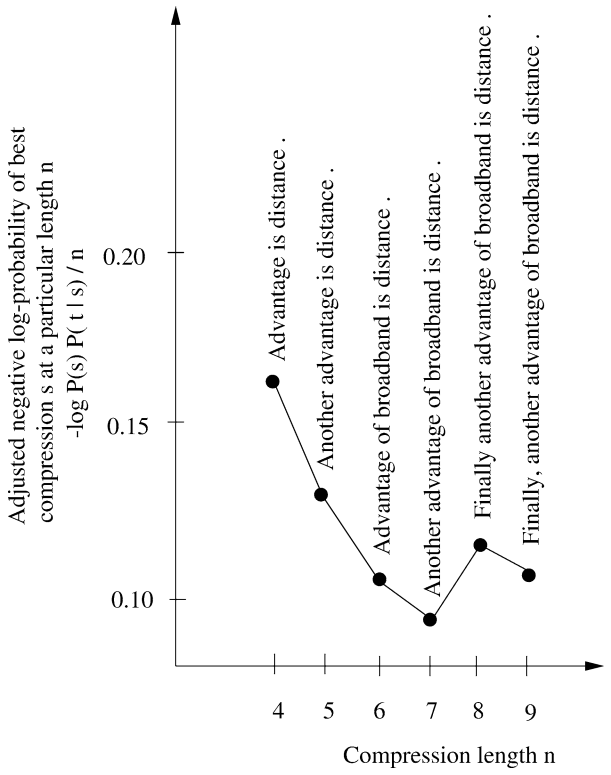

Fig. 4. Adjusted log-probabilities for top-scoring compressions at various lengths (lower is better).

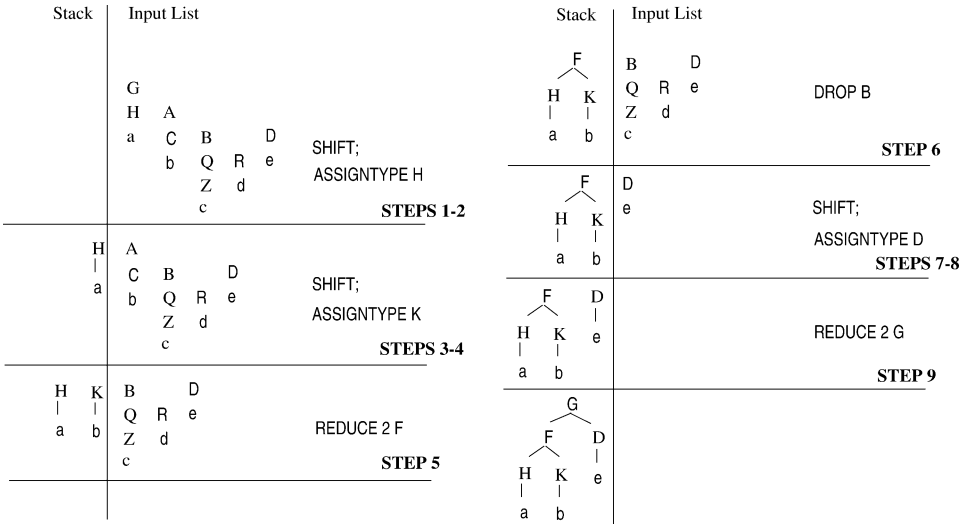

Fig. 5. Example of incremental tree compression.

- SHIFT operations transfer the first word from the input list into the stack.
- REDUCE operations pop the k syntactic trees located at the top of the stack; combine them into a new tree; and push the new tree on the top of the stack. Reduce operations are used to derive the structure of the syntactic tree of the short sentence.
- DROP operations are used to delete from the input list subsequences of words that correspond to syntactic constituents. A DROP X operations deletes from the input list all words that are spanned by constituent X in t.
- ASSIGNTYPE operations are used to change the label of trees at the top of the stack. These actions assign POS tags to the words in the compressed sentence, which may be different from the POS tags in the original sentence.

The decision-based model is more flexible than the channel model because it enables the derivation of trees whose skeleton can differ quite drastically from that of the tree given as input. For example, using the channel model, we are unable to obtain tree $s2$ from t. However, the four operations listed above enable us to rewrite a tree t into *any* tree s, as long as an in-order traversal of the leaves of s produces a sequence of words that occur in the same order as the words in the tree t.[1] For example, the tree $s2$ can be obtained from tree t by following this sequence of actions, whose effects are shown in Fig. 5: SHIFT; ASSIGNTYPE H; SHIFT; ASSIGNTYPE K; REDUCE 2 F; DROP B; SHIFT; ASSIGNTYPE D; REDUCE 2 G.

To save space, we show SHIFT and ASSIGNTYPE operations on the same line; however, the reader should understand that they correspond to two distinct actions. As one can see, the ASSIGNTYPE K operation rewrites the POS tag of the word b; the REDUCE operations modify the skeleton of the tree given as input. To increase readability, the input list is shown in a format that resembles as closely as possible the graphical representation of the trees in Fig. 1.

3.1. Learning the parameters of the decision-based model

We associate with each configuration of our shift-reduce-drop, rewriting model a learning case. The cases are generated automatically by a program that derives sequences of actions that map each of the large trees in our corpus into smaller trees. The rewriting procedure simulates a bottom-up reconstruction of the smaller trees.

Overall, the 1067 pairs of long and short sentences yielded 46383 learning cases. Each case was labeled with one action name from a set of 210 possible actions: There are 37 distinct ASSIGNTYPE actions, one for each POS tag. There are 63 distinct DROP actions, one for each type of syntactic constituent that can be deleted during compression. There are 109 distinct REDUCE actions, one for each type of reduce operation that is applied during the reconstruction of the compressed sentence. And there is one SHIFT operation. Given a tree t and an arbitrary configuration of the stack and input list, the purpose of the decision-based classifier is to learn what action to choose from the set of 210 possible actions.

[1] Marcu et al. [24] discuss a decision-based model of tree rewriting that permits leaves to be re-ordered as well.

To each learning example, we associated a set of 99 features from the following two classes:

Operational features reflect the number of trees in the stack, the input list, and the types of the last five operations. They also encode information that denote the syntactic category of the root nodes of the partial trees built up to a certain time. Examples of such features are: numberTreesInStack, wasPreviousOperationShift, syntacticLabelOfTreeAtTheTopOfStack, etc.

Original-tree-specific features denote the syntactic constituents that start with the first unit in the input list. Examples of such features are: inputListStartsWithA_CC, inputListStartsWithA_PP, etc.

The decision-based compression module uses the C4.5 program [27] in order to learn decision trees that specify how large syntactic trees can be compressed into shorter trees. A ten-fold cross-validation evaluation of the action classifier yielded an accuracy of 87.16% ($\pm$ 0.14), i.e., 87.16% of the time the system is faced with an action choice (SHIFT, REDUCE, ...), it makes the same choice as in the human-generated data.

A majority baseline classifier that chooses the action SHIFT has an accuracy of 28.72%.

3.1.1. Examples of learned compression rules

Given our training data, the decision-based classifier learned automatically rules such as those shown in Fig. 6. Rule 1 enables the deletion of WH prepositional phrases in the context in which they follow other constituents that the program decided to delete. Rule 2 enables the deletion of WHNP constituents. Since this deletion is carried out only when the stack contains only one NP constituent, it follows that this rule is applied only in conjunction with complex nounphrases that occur at the beginning of sentences. Rule 3 enables the deletion of adjectival phrases.

Rule 1: IF previous operation was not "Reduce" AND
previous operation was not "Shift" AND
previous operation was not "AssignType" AND
the input list starts with a syntactic constituent of type WHPP
THEN drop from the input list the words subsumed by WHPP.

Rule 2: IF there is only one tree in the stack AND
previous operation was "Reduce" AND
the syntactic label of the tree in the stack is NP-A AND
the input list starts with a syntactic constituent of type WHNP
THEN drop from the input list the words subsumed by WHNP.

Rule 3: IF previous operation was "Drop" AND
the input list starts with a syntactic constituent of type ADJP AND
the input list does not start with a syntactic constituent of type NP
THEN drop from the input list the words subsumed by ADJP.

Fig. 6. Examples of rules that were learned automatically by the c4.5 program.

3.2. Employing the decision-based model

To compress sentences, we apply the shift-reduce-drop model in a deterministic fashion. We parse the sentence to be compressed [12] and we initialize the input list with the words in the sentence and the syntactic constituents that "begin" at each word, as shown in Fig. 5. We then incrementally inquire the learned classifier what action to perform, and we simulate the execution of that action. The procedure ends when the input list is empty and when the stack contains only one tree. An inorder traversal of the leaves of this tree produces the compressed version of the sentence given as input.

Since the model is deterministic, it produces only one output. The advantage is that the compression is very fast: it takes only a few milliseconds per sentence, not counting parsing. The disadvantage is that it does not produce a range of compressions, from which another system may subsequently choose. It is straightforward though to extend the model within a probabilistic framework by applying, for example, the techniques used by [20].

4. Evaluation

To evaluate our compression algorithms, we randomly selected 32 sentence pairs from our parallel corpus, which we will refer to as the *Test Corpus*. We used the other 1035 sentence pairs for training. Fig. 7 shows three sentences from the Test Corpus, together with the compressions produced by humans, our compression algorithms, and a baseline algorithm that produces compressions with highest word-bigram scores. The examples are

Original:	Beyond the basic level, the operations of the three products vary widely.
Baseline:	Beyond the basic level, the operations of the three products vary widely.
Noisy-channel:	The operations of the three products vary widely.
Decision-based:	The operations of the three products vary widely.
Humans:	The operations of the three products vary widely.
Original:	Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.
Baseline:	Arborscan and worked in, but it very large dxf.
Noisy-channel:	Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.
Decision-based:	Arborscan is reliable and worked accurately in testing very large dxf files.
Humans:	Arborscan produces very large dxf files.
Original:	Many debugging features, including user-defined break points and variable-watching and message-watching windows, have been added.
Baseline:	Debugging, user-defined and variable-watching and message-watching, have been.
Noisy-channel:	Many debugging features, including user-defined points and variable-watching and message-watching windows, have been added.
Decision-based:	Many debugging features.
Humans:	Many debugging features have been added.

Fig. 7. Selected compression examples.

chosen so as to reflect good, average, and bad performance cases. The first sentence is compressed in the same manner by humans and our algorithms (the baseline algorithm chooses though not to compress this sentence). For the second example, the output of the Decision-based algorithm is grammatical, but the semantics is negatively affected. The noisy-channel algorithm deletes only the word "break", which affects the correctness of the output less. In the last example, the noisy-channel model is again more conservative and decides not to drop any constituents. In contrast, the decision-based algorithm compresses the input substantially, but it fails to produce a grammatical output.

We presented each original sentence in the *Test Corpus* to four judges, together with four compressions of it: the human generated compression, the outputs of the noisy-channel and decision-based algorithms, and the output of the baseline algorithm. The judges were told that all outputs were generated automatically. The order of the outputs was scrambled randomly across test cases.

To avoid confounding, the judges participated in two experiments. In the first experiment, they were asked to determine on a scale from 1 to 5 how well the systems did with respect to selecting the most important words in the original sentence. In the second experiment, they were asked to determine on a scale from 1 to 5 how grammatical the outputs were.

We also investigated how sensitive our algorithms are with respect to the training data by carrying out the same experiments on sentences of a different genre. To this end, we took the first sentence of the first 26 articles made available in 1999 on the scientific *cmplg* archive. We created a second test corpus, which we will refer to as the *Cmplg Corpus*, by generating by ourselves compressed grammatical versions of these sentences. (Training was done on the Ziff–Davis corpus, as before.) Since some of the sentences in this corpus were extremely long, the baseline algorithm could not produce compressed versions.

The results in Table 1 show compression rates, and mean and standard deviation results across all judges, for each algorithm and corpus. The results show that the decision-based algorithm is the most aggressive: on average, it compresses sentences to about half of their original size. The compressed sentences produced by both algorithms are more "grammatical" and contain more important words than the sentences produced by the baseline. T-test experiments showed these differences to be statistically significant at $p < 0.01$ both for individual judges and for average scores across all judges. T-tests showed no significant statistical differences between the two algorithms. As Table 1 shows, the performance of

Table 1
Experimental results

Corpus	Avg. orig. sent. length		Baseline	Noisy-channel	Decision-based	Humans
Test	21 words	Compression	63.70%	70.37%	57.19%	53.33%
		Grammaticality	1.78 ± 1.19	4.34 ± 1.02	4.30 ± 1.33	4.92 ± 0.18
		Importance	2.17 ± 0.89	3.38 ± 0.67	3.54 ± 1.00	4.24 ± 0.52
Cmplg	26 words	Compression	–	65.68%	54.25%	65.68%
		Grammaticality	–	4.22 ± 0.99	3.72 ± 1.53	4.97 ± 0.08
		Importance	–	3.42 ± 0.97	3.24 ± 0.68	4.32 ± 0.54

the compression algorithms is much closer to human performance than baseline performance; yet, humans perform statistically better than our algorithms at $p < 0.01$.

When applied to sentences of a different genre, the performance of the noisy-channel compression algorithm degrades smoothly, while the performance of the decision-based algorithm drops sharply. This is due to a few sentences in the *Cmplg Corpus* that the decision-based algorithm over-compressed to only two or three words. We suspect that this problem can be fixed if the decision-based compression module is extended in the style of Magerman [20], by computing probabilities across the sequences of decisions that correspond to a compressed sentence. Likewise, there are substantial gains to be had in noisy-channel modeling—we see clearly in the data many statistical dependencies and processes that are not captured in our simple initial models. More grammatical output will come from taking account of subcategory and head-modifier statistics (in addition to simple word-bigrams), and an expanded channel model will allow for more tree manipulation possibilities. Work on extending the algorithms presented in this paper to compressing multiple sentences is currently underway.

5. Conclusions

We have described sentence compression, a summarization task that requires reasoning about fluency and the relative importance of different pieces of text. We have presented corpus-based methods for attacking this problem, one using the noisy-channel framework, and other using a decision-based model. While previous corpus-based work in summarization has focused on keyword extraction, this work shows that it is feasible to construct new whole sentences by analyzing existing, manually produced, compressions.

In addition to having many useful applications, we view sentence compression as a stepping stone towards building high-quality, document-level summarization systems. We believe that corpus-based approaches of the kind described in this paper provide the best way to scale up to the full problem of text compression, as vast amounts of data are widely available in the form of document/abstract pairs. The next steps along this path will be to devise models of the human summarizer that fit this data and can be trained on it.

References

[1] M. Banko, V. Mittal, M. Witbrock, Headline generation based on statistical translation, in: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000, pp. 318–325.

[2] R. Barzilay, N. Elhadad, K. McKeown, Sentence ordering in multidocument summarization, in: Proceedings of the First International Conference on Human Language Technology Research (HLT-01), San Diego, CA, 2001, pp. 149–156.

[3] R. Barzilay, K. McKeown, M. Elhadad, Information fusion in the context of multi-document summarization, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), University of Maryland, 1999, pp. 550–557.

[4] A. Berger, J. Lafferty, Information retrieval as statistical translation, in: Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, CA, 1999, pp. 222–229.

[5] A. Berger, V. Mittal, Query-relevant summarization using FAQs, in: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000, pp. 294–301.

[6] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer, The mathematics of statistical machine translation: Parameter estimation, Comput. Linguistics 19 (2) (1993) 263–311.
[7] O. Buyukkokten, H. Garcia-Molina, A. Paepcke, Seeing the whole in parts: Text summarization for web browsing on handheld devices, in: Proceedings of the 10th International WWW Conference, Hong Kong, China, 2001.
[8] Y. Canning, J. Tait, J. Archibald, R. Crawley, Cohesive generation of syntactically simplified newspaper text, in: Workshop on Robust Methods in Analysis of Natural Language Data, Lausanne, 2000, pp. 145–150.
[9] J. Carroll, G. Minnen, Y. Canning, S. Devlin, J. Tait, Practical simplification of English newspaper text to assist aphasic readers, in: Proceedings of the AAAI-98 Workshop on Integrating AI and Assistive Technology, Madison, WI, 1998.
[10] R. Chandrasekar, C. Doran, B. Srinivas, Motivations and methods for text simplification, in: Proceedings of the International Conference on Computational Linguistics (COLING-96), Copenhagen, 1996.
[11] K. Church, A stochastic parts program and noun phrase parser for unrestricted text, in: Proceedings of the Second Conference on Applied Natural Language Processing, Austin, TX, 1988, pp. 136–143.
[12] M. Collins, Three generative lexicalized models for statistical parsing, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), Madrid, Spain, 1997, pp. 16–23.
[13] G. Grefenstette, Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind, in: Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization, Stanford University, CA, 1998, pp. 111–118.
[14] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, MA, 1997.
[15] H. Jing, K. McKeown, The decomposition of human-written summary sentences, in: Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, CA, 1999.
[16] R. Jing, A. Hauptmann, Title generation for machine-translated documents, in: Proceedings of IJCAI-01, Seattle, WA, 2001, pp. 1229–1234.
[17] K. Knight, J. Graehl, Machine transliteration, Comput. Linguistics 24 (4) (1998) 599–612.
[18] I. Langkilde, Forest-based statistical sentence generation, in: Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, 2000.
[19] N. Linke-Ellis, Closed captioning in America: Looking beyond compliance, in: Proceedings of the TAO Workshop on TV Closed Captions for the Hearing Impaired People, Tokyo, Japan, 1999, pp. 43–59.
[20] D. Magerman, Statistical decision-tree models for parsing, in: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, 1995, pp. 276–283.
[21] K. Mahesh, Hypertext summary extraction for fast document browsing, in: Proceedings of AAAI Spring Symposium on Natural Language Processing for the World Wide Web, Stanford, CA, 1997, pp. 95–104.
[22] I. Mani, B. Gates, E. Bloedorn, Improving summaries by revising them, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, MD, 1999, pp. 558–565.
[23] I. Mani, M. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999.
[24] D. Marcu, L. Carlson, M. Watanabe, The automatic translation of discourse structures, in: Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics NAACL-2000, Seattle, WA, 2000, pp. 9–17.
[25] D. Marcu, L. Gerber, An inquiry into the nature of multidocument abstract extracts and their evaluation, in: Proceedings of the NAACL-01 Workshop on Text Summarization, Pittsburgh, PA, 2001.
[26] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, E. Eskin, Towards multidocument summarization by reformulation: Progress and prospects, in: Proceedings of AAAI-99, Orlando, FL, 1999, pp. 453–460.
[27] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[28] J. Robert-Ribes, S. Pfeiffer, R. Ellison, D. Burnham, Semi-automatic captioning of TV programs, an Australian perspective, in: Proceedings of the TAO Workshop on TV Closed Captions for the Hearing Impaired People, Tokyo, Japan, 1999, pp. 87–100.
[29] M. Witbrock, V. Mittal, Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries, in: Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99), Poster Session, Berkeley, CA, 1999, pp. 315–316.
[30] J. Zelle, R. Mooney, Learning to parse database queries using inductive logic programming, in: Proceedings AAAI-96, Portland, OR, 1996, pp. 1050–1055.