Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures

Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures

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

Abstract

Sentence compression is the task of producing a summary at the sentence level. This paper focuses on three aspects of this task which have not received detailed treatment in the literature: training requirements, scalability, and automatic evaluation. We provide a novel comparison between a supervised constituent-based and an weakly supervised word-based compression algorithm and examine how these models port to different domains (written vs. spoken text). To achieve this, a human-authored compression corpus has been created and our study highlights potential problems with the automatically gathered compression corpora currently used. Finally, we assess whether automatic evaluation measures can be used to determine compression quality.

1 Introduction

Automatic sentence compression has recently attracted much attention, in part because of its affinity with summarisation. The task can be viewed as producing a summary of a single sentence that retains the most important information while remaining grammatically correct. An ideal compression algorithm will involve complex text rewriting operations such as word reordering, paraphrasing, substitution, deletion, and insertion. In default of a more sophisticated compression algorithm, current approaches have simplified the problem to a single rewriting operation, namely word deletion. More formally, given an input sentence of words $W = w_1, w_2, \ldots, w_n$, a compression is formed by dropping any subset of these words. Viewing the task as word removal reduces the number of possible compressions to 2^n; naturally, many of these compressions will not be reasonable or grammatical (Knight and Marcu 2002).

Sentence compression could be usefully employed in wide range of applications. For example, to automatically generate subtitles for television programs; the transcripts cannot usually be used verbatim due to the rate of speech being too high (Vandeghinste and Pan 2004). Other applications include compressing text to be displayed on small screens (Corston-Oliver 2001) such as mobile phones or PDAs, and producing audio scanning devices for the blind (Grefenstette 1998).

Algorithms for sentence compression fall into two broad classes depending on their training requirements. Many algorithms exploit parallel corpora (Jing 2000; Knight and Marcu 2002; Riezler et al. 2003; Nguyen et al. 2004a; Turner and Charniak 2005; McDonald 2006) to learn the correspondences between long and short sentences in a supervised manner, typically using a rich feature space induced from parse trees. The learnt rules effectively describe which constituents should be deleted in a given context. Approaches that do not employ parallel corpora require minimal or no supervision. They operationalise compression in terms of word deletion without learning specific rules and can therefore rely on little linguistic knowledge such as part-of-speech tags or merely the lexical items alone (Hori and Furui 2004). Alternatively, the rules of compression are approximated from a non-parallel corpus (e.g., the Penn Treebank) by considering context-free grammar derivations with matching expansions (Turner and Charniak 2005).

Previous approaches have been developed and tested almost exclusively on written text, a notable exception being Hori and Furui (2004) who focus on spoken language. While parallel corpora of original-compressed sentences are not naturally available in the way multilingual corpora are, researchers have obtained such corpora automatically by exploiting documents accompanied by abstracts. Automatic corpus creation affords the opportunity to study compression mechanisms

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 377–384,
Sydney, July 2006.

cheaply, yet these mechanisms may not be representative of human performance. It is unlikely that authors routinely carry out sentence compression while creating abstracts for their articles. Collecting human judgements is the method of choice for evaluating sentence compression models. However, human evaluations tend to be expensive and cannot be repeated frequently; furthermore, comparisons across different studies can be difficult, particularly if subjects employ different scales, or are given different instructions.

In this paper we examine some aspects of the sentence compression task that have received little attention in the literature. First, we provide a novel comparison of supervised and weakly supervised approaches. Specifically, we study how constituent-based and word-based methods port to different domains and show that the latter tend to be more robust. Second, we create a corpus of human-authored compressions, and discuss some potential problems with currently used compression corpora. Finally, we present automatic evaluation measures for sentence compression and examine whether they correlate reliably with behavioural data.

2 Algorithms for Sentence Compression

In this section we give a brief overview of the algorithms we employed in our comparative study. We focus on two representative methods, Knight and Marcu's (2002) decision-based model and Hori and Furui's (2004) word-based model.

The decision-tree model operates over parallel corpora and offers an intuitive formulation of sentence compression in terms of tree rewriting. It has inspired many discriminative approaches to the compression task (Riezler et al. 2003; Nguyen et al. 2004b; McDonald 2006) and has been extended to languages other than English (see Nguyen et al. 2004a). We opted for the decision-tree model instead of the also well-known noisy-channel model (Knight and Marcu 2002; Turner and Charniak 2005). Although both models yield comparable performance, Turner and Charniak (2005) show that the latter is not an appropriate compression model since it favours uncompressed sentences over compressed ones.[1]

Hori and Furui's (2004) model was originally developed for Japanese with spoken text in mind, it requires minimal supervision, and little linguistic knowledge. It therefor holds promise for languages and domains for which text processing tools (e.g., taggers, parsers) are not readily available. Furthermore, to our knowledge, its performance on written text has not been assessed.

SHIFT transfers the first word from the input list onto the stack.
REDUCE pops the syntactic trees located at the top of the stack, combines them into a new tree and then pushes the new tree onto the top of the stack.
DROP deletes from the input list subsequences of words that correspond to a syntactic constituent.
ASSIGNTYPE changes the label of the trees at the top of the stack (i.e., the POS tag of words).

Table 1: Stack rewriting operations

2.1 Decision-based Sentence Compression

In the decision-based model, sentence compression is treated as a deterministic rewriting process of converting a long parse tree, *l*, into a shorter parse tree *s*. The rewriting process is decomposed into a sequence of shift-reduce-drop actions that follow an extended shift-reduce parsing paradigm.

The compression process starts with an empty stack and an input list that is built from the original sentence's parse tree. Words in the input list are labelled with the name of all the syntactic constituents in the original sentence that start with it. Each stage of the rewriting process is an operation that aims to reconstruct the compressed tree. There are four types of operations that can be performed on the stack, they are illustrated in Table 1.

Learning cases are automatically generated from a parallel corpus. Each learning case is expressed by a set of features and represents one of the four possible operations for a given stack and input list. Using the C4.5 program (Quinlan 1993) a decision-tree model is automatically learnt. The model is applied to a parsed original sentence in a deterministic fashion. Features for the current state of the input list and stack are extracted and the classifier is queried for the next operation to perform. This is repeated until the input list is empty and the stack contains only one item (this corresponds to the parse for the compressed tree). The compressed sentence is recovered by traversing the leaves of the tree in order.

2.2 Word-based Sentence Compression

The decision-based method relies exclusively on parallel corpora; the caveat here is that appropriate training data may be scarce when porting this model to different text domains (where abstracts

[1] The noisy-channel model uses a source model trained on uncompressed sentences. This means that the most likely compressed sentence will be identical to the original sentence as the likelihood of a constituent deletion is typically far lower than that of leaving it in.

are not available for automatic corpus creation) or languages. To alleviate the problems inherent with using a parallel corpus, we have modified a weakly supervised algorithm originally proposed by Hori and Furui (2004). Their method is based on word deletion; given a prespecified compression length, a compression is formed by preserving the words which maximise a scoring function.

To make Hori and Furui's (2004) algorithm more comparable to the decision-based model, we have eliminated the compression length parameter. Instead, we search over all lengths to find the compression that gives the maximum score. This process yields more natural compressions with varying lengths. The original score measures the significance of each word (I) in the compression and the linguistic likelihood (L) of the resulting word combinations.[2] We add some linguistic knowledge to this formulation through a function (SOV) that captures information about subjects, objects and verbs. The compression score is given in Equation (1). The lambdas (λ_I, λ_{SOV}, λ_L) weight the contribution of the individual scores:

$$\begin{aligned} S(V) &= \sum_{i=1}^{M} \lambda_I I(v_i) + \lambda_{sov} SOV(v_i) \\ &\quad + \lambda_L L(v_i|v_{i-1}, v_{i-2}) \end{aligned} \tag{1}$$

The sentence $V = v_1, v_2, \ldots, v_m$ (of M words) that maximises the score $S(V)$ is the best compression for an original sentence consisting of N words ($M < N$). The best compression can be found using dynamic programming. The λ's in Equation (1) can be either optimised using a small amount of training data or set manually (e.g., if short compressions are preferred to longer ones, then the language model should be given a higher weight). Alternatively, weighting could be dispensed with by including a normalising factor in the language model. Here, we follow Hori and Furui's (2004) original formulation and leave the normalisation to future work. We next introduce each measure individually.

Word significance score The word significance score I measures the relative importance of a word in a document. It is similar to tf-idf, a term weighting score commonly used in information retrieval:

$$I(w_i) = f_i \log \frac{F_A}{F_i} \tag{2}$$

Where w_i is the topic word of interest (topic words are either nouns or verbs), f_i is the frequency of w_i in the document, F_i is the corpus frequency of w_i and F_A is the sum of all topic word occurrences in the corpus ($\sum_i F_i$).

Linguistic score The linguistic score's $L(v_i|v_{i-1}, v_{i-2})$ responsibility is to select some function words, thus ensuring that compressions remain grammatical. It also controls which topic words can be placed together. The score measures the n-gram probability of the compressed sentence.

SOV Score The SOV score is based on the intuition that subjects, objects and verbs should not be dropped while words in other syntactic roles can be considered for removal. This score is based solely on the contents of the sentence considered for compression without taking into account the distribution of subjects, objects or verbs, across documents. It is defined in (3) where f_i is the document frequency of a verb, or word bearing the subject/object role and $\lambda_{default}$ is a constant weight assigned to all other words.

$$SOV(w_i) = \begin{cases} f_i & \text{if } w_i \text{ in subject, object} \\ & \text{or verb role} \\ \lambda_{default} & \text{otherwise} \end{cases} \tag{3}$$

The SOV score is only applied to the head word of subjects and objects.

3 Corpora

Our intent was to assess the performance of the two models just described on written and spoken text. The appeal of written text is understandable since most summarisation work today focuses on this domain. Speech data not only provides a natural test-bed for compression applications (e.g., subtitle generation) but also poses additional challenges. Spoken utterances can be ungrammatical, incomplete, and often contain artefacts such as false starts, interjections, hesitations, and disfluencies. Rather than focusing on spontaneous speech which is abundant in these artefacts, we conduct our study on the less ambitious domain of broadcast news transcripts. This lies inbetween the extremes of written text and spontaneous speech as it has been scripted beforehand and is usually read off an autocue.

One stumbling block to performing a comparative study between written data and speech data is that there are no naturally occurring parallel

[2] Hori and Furui (2004) also have a confidence score based upon how reliable the output of an automatic speech recognition system is. However, we need not consider this score when working with written text and manual transcripts.

speech corpora for studying compression. Automatic corpus creation is not a viable option either, speakers do not normally create summaries of their own utterances. We thus gathered our own corpus by asking humans to generate compressions for speech transcripts.

In what follows we describe how the manual compressions were performed. We also briefly present the written corpus we used for our experiments. The latter was automatically constructed and offers an interesting point of comparison with our manually created corpus.

Broadcast News Corpus Three annotators were asked to compress 50 broadcast news stories (1,370 sentences) taken from the HUB-4 1996 English Broadcast News corpus provided by the LDC. The HUB-4 corpus contains broadcast news from a variety of networks (CNN, ABC, CSPAN and NPR) which have been manually transcribed and split at the story and sentence level. Each document contains 27 sentences on average and the whole corpus consists of 26,151 tokens.[3] The Robust Accurate Statistical Parsing (RASP) toolkit (Briscoe and Carroll 2002) was used to automatically tokenise the corpus.

Each annotator was asked to perform sentence compression by removing tokens from the original transcript. Annotators were asked to remove words while: (a) preserving the most important information in the original sentence, and (b) ensuring the compressed sentence remained grammatical. If they wished they could leave a sentence uncompressed by marking it as inappropriate for compression. They were not allowed to delete whole sentences even if they believed they contained no information content with respect to the story as this would blur the task with abstracting.

Ziff-Davis Corpus Most previous work (Jing 2000; Knight and Marcu 2002; Riezler et al. 2003; Nguyen et al. 2004a; Turner and Charniak 2005; McDonald 2006) has relied on automatically constructed parallel corpora for training and evaluation purposes. The most popular compression corpus originates from the Ziff-Davis corpus — a collection of news articles on computer products. The corpus was created by matching sentences that occur in an article with sentences that occur in an abstract (Knight and Marcu 2002). The abstract sentences had to contain a subset of the original sentence's words and the word order had to remain the same.

[3]The compression corpus is available at `http://homepages.inf.ed.ac.uk/s0460084/data/`.

	A1	A2	A3	Av.	Ziff-Davis
Comp%	88.0	79.0	87.0	84.4	97.0
CompR	73.1	79.0	70.0	73.0	47.0

Table 2: Compression Rates (Comp% measures the percentage of sentences compressed; CompR is the mean compression rate of all sentences)

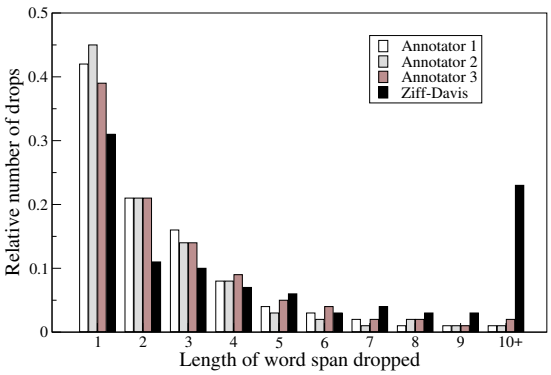

Figure 1: Distribution of span of words dropped

Comparisons Following the classification scheme adopted in the British National Corpus (Burnard 2000), we assume throughout this paper that Broadcast News and Ziff-Davis belong to different domains (spoken vs. written text) whereas they represent the same genre (i.e., news). Table 2 shows the percentage of sentences which were compressed (Comp%) and the mean compression rate (CompR) for the two corpora. The annotators compress the Broadcast News corpus to a similar degree. In contrast, the Ziff-Davis corpus is compressed much more aggressively with a compression rate of 47%, compared to 73% for Broadcast News. This suggests that the Ziff-Davis corpus may not be a true reflection of human compression performance and that humans tend to compress sentences more conservatively than the compressions found in abstracts.

We also examined whether the two corpora differ with regard to the length of word spans being removed. Figure 1 shows how frequently word spans of varying lengths are being dropped. As can be seen, a higher percentage of long spans (five or more words) are dropped in the Ziff-Davis corpus. This suggests that the annotators are removing words rather than syntactic constituents, which provides support for a model that can act on the word level. There is no statistically significant difference between the length of spans dropped between the annotators, whereas there is a significant difference ($p < 0.01$) between the annotators' spans and the Ziff-Davis' spans (using the

Wilcoxon Test).

The compressions produced for the Broadcast News corpus may differ slightly to the Ziff-Davis corpus. Our annotators were asked to perform sentence compression explicitly as an isolated task rather than indirectly (and possibly subconsciously) as part of the broader task of abstracting, which we can assume is the case with the Ziff-Davis corpus.

4 Automatic Evaluation Measures

Previous studies relied almost exclusively on human judgements for assessing the well-formedness of automatically derived compressions. Although human evaluations of compression systems are not as large-scale as in other fields (e.g., machine translation), they are typically performed once, at the end of the development cycle. Automatic evaluation measures would allow more extensive parameter tuning and crucially experimentation with larger data sets. Most human studies to date are conducted on a small compression sample, the test portion of the Ziff-Davis corpus (32 sentences). Larger sample sizes would expectedly render human evaluations time consuming and generally more difficult to conduct frequently. Here, we review two automatic evaluation measures that hold promise for the compression task.

Simple String Accuracy (SSA, Bangalore et al. 2000) has been proposed as a baseline evaluation metric for natural language generation. It is based on the string edit distance between the generated output and a gold standard. It is a measure of the number of insertion (I), deletion (D) and substitution (S) errors between two strings. It is defined in (4) where R is the length of the gold standard string.

$$\text{Simple String Accuracy} = (1 - \frac{I+D+S}{R}) \quad (4)$$

The SSA score will assess whether appropriate words have been included in the compression.

Another stricter automatic evaluation method is to compare the grammatical relations found in the system compressions against those found in a gold standard. This allows us "to measure the semantic aspects of summarisation quality in terms of grammatical-functional information" (Riezler et al. 2003). The standard metrics of precision, recall and F-score can then be used to measure the quality of a system against a gold standard. Our implementation of the F-score measure used the grammatical relations annotations provided by RASP (Briscoe and Carroll 2002). This parser is particularly appropriate for the compression task since it provides parses for both full sentences and sentence fragments and is generally robust enough to analyse semi-grammatical compressions. We calculated F-score over all the relations provided by RASP (e.g., subject, direct/indirect object, modifier; 15 in total).

Correlation with human judgements is an important prerequisite for the wider use of automatic evaluation measures. In the following section we describe an evaluation study examining whether the measures just presented indeed correlate with human ratings of compression quality.

5 Experimental Set-up

In this section we present our experimental set-up for assessing the performance of the two algorithms discussed above. We explain how different model parameters were estimated. We also describe a judgement elicitation study on automatic and human-authored compressions.

Parameter Estimation We created two variants of the decision-tree model, one trained on the Ziff-Davis corpus and one on the Broadcast News corpus. We used 1,035 sentences from the Ziff-Davis corpus for training; the same sentences were previously used in related work (Knight and Marcu 2002). The second variant was trained on 1,237 sentences from the Broadcast News corpus. The training data for both models was parsed using Charniak's (2000) parser. Learning cases were automatically generated using a set of 90 features similar to Knight and Marcu (2002).

For the word-based method, we randomly selected 50 sentences from each training set to optimise the lambda weighting parameters[4]. Optimisation was performed using Powell's method (Press et al. 1992). Recall from Section 2.2 that the compression score has three main parameters: the significance, linguistic, and *SOV* scores. The significance score was calculated using 25 million tokens from the Broadcast News corpus (spoken variant) and 25 million tokens from the North American News Text Corpus (written variant). The linguistic score was estimated using a trigram language model. The language model was trained on the North Ameri-

[4]To treat both models on an equal footing, we attempted to train the decision-tree model solely on 50 sentences. However, it was unable to produce any reasonable compressions, presumably due to insufficient learning instances.

can corpus (25 million tokens) using the CMU-Cambridge Language Modeling Toolkit (Clarkson and Rosenfeld 1997) with a vocabulary size of 50,000 tokens and Good-Turing discounting. Subjects, objects, and verbs for the *SOV* score were obtained from RASP (Briscoe and Carroll 2002).

All our experiments were conducted on sentences for which we obtained syntactic analyses. RASP failed on 17 sentences from the Broadcast news corpus and 33 from the Ziff-Davis corpus; Charniak's (2000) parser successfully parsed the Broadcast News corpus but failed on three sentences from the Ziff-Davis corpus.

Evaluation Data We randomly selected 40 sentences for evaluation purposes, 20 from the testing portion of the Ziff-Davis corpus (32 sentences) and 20 sentences from the Broadcast News corpus (133 sentences were set aside for testing). This is comparable to previous studies which have used the 32 test sentences from the Ziff-Davis corpus. None of the 20 Broadcast News sentences were used for optimisation. We ran the decision-tree system and the word-based system on these 40 sentences. One annotator was randomly selected to act as the gold standard for the Broadcast News corpus; the gold standard for the Ziff-Davis corpus was the sentence that occurred in the abstract. For each original sentence we had three compressions; two generated automatically by our systems and a human authored gold standard. Thus, the total number of compressions was 120 (3x40).

Human Evaluation The 120 compressions were rated by human subjects. Their judgements were also used to examine whether the automatic evaluation measures discussed in Section 4 correlate reliably with behavioural data. Sixty unpaid volunteers participated in our elicitation study, all were self reported native English speakers. The study was conducted remotely over the Internet. Participants were presented with a set of instructions that explained the task and defined sentence compression with the aid of examples. They first read the original sentence with the compression hidden. Then the compression was revealed by pressing a button. Each participant saw 40 compressions. A Latin square design prevented subjects from seeing two different compressions of the same sentence. The order of the sentences was randomised. Participants were asked to rate each compression they saw on a five point scale taking into account the information retained by the compression and its grammaticality. They were told all compressions were automatically generated. Examples of the compressions our participants saw are given in Table 3.

o:	Apparently Fergie very much wants to have a career in television.
d:	A career in television.
w:	Fergie wants to have a career in television.
g:	Fergie wants a career in television.
o:	Many debugging features, including user-defined break points and variable-watching and message-watching windows, have been added.
d:	Many debugging features.
w:	Debugging features, and windows, have been added.
g:	Many debugging features have been added.
o:	As you said, the president has just left for a busy three days of speeches and fundraising in Nevada, California and New Mexico.
d:	As you said, the president has just left for a busy three days.
w:	You said, the president has left for three days of speeches and fundraising in Nevada, California and New Mexico.
g:	The president left for three days of speeches and fundraising in Nevada, California and New Mexico.

Table 3: Compression examples (o: original sentence, d: decision-tree compression, w: word-based compression, g: gold standard)

6 Results

Our experiments were designed to answer three questions: (1) Is there a significant difference between the compressions produced by supervised (constituent-based) and weakly unsupervised (word-based) approaches? (2) How well do the two models port across domains (written vs. spoken text) and corpora types (human vs. automatically created)? (3) Do automatic evaluation measures correlate with human judgements?

One of our first findings is that the the decision-tree model is rather sensitive to the style of training data. The model cannot capture and generalise single word drops as effectively as constituent drops. When the decision-tree is trained on the Broadcast News corpus, it is unable to create suitable compressions. On the evaluation data set, 75% of the compressions produced are the original sentence or the original sentence with one word removed. It is possible that the Broadcast News compression corpus contains more varied compressions than those of the Ziff-Davis and therefore a larger amount of training data would be required to learn a reliable decision-tree model. We thus used the Ziff-Davis trained decision-tree model to obtain compressions for both corpora.

Our results are summarised in Tables 4 and 5. Table 4 lists the average compression rates for

Broadcast News	CompR	SSA	F-score
Decision-tree	0.55	0.34	0.40
Word-based	0.72	0.51	0.54
gold standard	0.71	–	–

Ziff-Davis	CompR	SSA	F-score
Decision-tree	0.58	0.20	0.34
Word-based	0.60	0.19	0.39
gold standard	0.54	–	–

Table 4: Results using automatic evaluation measures

Compression	Broadcast News	Ziff-Davis
Decision-tree	2.04	2.34
Word-based	2.78	2.43
gold standard	3.87	3.53

Table 5: Mean ratings from human evaluation

each model as well as the models' performance according to the two automatic evaluation measures discussed in Section 4. The row 'gold standard' displays human-produced compression rates. Table 5 shows the results of our judgement elicitation study.

The compression rates (CompR, Table 4) indicate that the decision-tree model compresses more aggressively than the word-based model. This is due to the fact that it mostly removes entire constituents rather than individual words. The word-based model is closer to the human compression rate. According to our automatic evaluation measures, the decision-tree model is significantly worse than the word-based model (using the Student t test, SSA $p < 0.05$, F-score $p < 0.05$) on the Broadcast News corpus. Both models are significantly worse than humans (SSA $p < 0.05$, F-score $p < 0.01$). There is no significant difference between the two systems using the Ziff-Davis corpus on both simple string accuracy and relation F-score, whereas humans significantly outperform the two systems.

We have performed an Analysis of Variance (ANOVA) to examine whether similar results are obtained when using human judgements. Statistical tests were done using the mean of the ratings (see Table 5). The ANOVA revealed a reliable effect of compression type by subjects and by items ($p < 0.01$). Post-hoc Tukey tests confirmed that the word-based model outperforms the decision-tree model ($\alpha < 0.05$) on the Broadcast news corpus; however, the two models are not significantly different when using the Ziff-Davis corpus. Both systems perform significantly worse than the gold standard ($\alpha < 0.05$).

Measure	Ziff-Davis	Broadcast News
SSA	0.171	0.348*
F-score	0.575**	0.532**
$^*p < 0.05$ $^{**}p < 0.01$		

Table 6: Correlation (Pearson's r) between evaluation measures and human ratings. Stars indicate level of statistical significance.

We next examine the degree to which the automatic evaluation measures correlate with human ratings. Table 6 shows the results of correlating the simple string accuracy (SSA) and relation F-score against compression judgements. The SSA does not correlate on both corpora with human judgements; it thus seems to be an unreliable measure of compression performance. However, the F-score correlates significantly with human ratings, yielding a correlation coefficient of $r = 0.575$ on the Ziff-Davis corpus and $r = 0.532$ on the Broadcast news. To get a feeling for the difficulty of the task, we assessed how well our participants agreed in their ratings using leave-one-out resampling (Weiss and Kulikowski 1991). The technique correlates the ratings of each participant with the mean ratings of all the other participants. The average agreement is $r = 0.679$ on the Ziff-Davis corpus and $r = 0.746$ on the Broadcast News corpus. This result indicates that F-score's agreement with the human data is not far from the human upper bound.

7 Conclusions and Future Work

In this paper we have provided a comparison between a supervised (constituent-based) and a minimally supervised (word-based) approach to sentence compression. Our results demonstrate that the word-based model performs equally well on spoken and written text. Since it does not rely heavily on training data, it can be easily extended to languages or domains for which parallel compression corpora are scarce. When no parallel corpora are available the parameters can be manually tuned to produce compressions. In contrast, the supervised decision-tree model is not particularly robust on spoken text, it is sensitive to the nature of the training data, and did not produce adequate compressions when trained on the human-authored Broadcast News corpus. A comparison of the automatically gathered Ziff-Davis corpus

with the Broadcast News corpus revealed important differences between the two corpora and thus suggests that automatically created corpora may not reflect human compression performance.

We have also assessed whether automatic evaluation measures can be used for the compression task. Our results show that grammatical relations-based F-score (Riezler et al. 2003) correlates reliably with human judgements and could thus be used to measure compression performance automatically. For example, it could be used to assess progress during system development or for comparing across different systems and system configurations with much larger test sets than currently employed.

In its current formulation, the only function driving compression in the word-based model is the language model. The word significance and *SOV* scores are designed to single out important words that the model should not drop. We have not yet considered any functions that encourage compression. Ideally these functions should be inspired from the underlying compression process. Finding such a mechanism is an avenue of future work. We would also like to enhance the word-based model with more linguistic knowledge; we plan to experiment with syntax-based language models and more richly annotated corpora.

Another important future direction lies in applying the unsupervised model presented here to languages with more flexible word order and richer morphology than English (e.g., German, Czech). We suspect that these languages will prove challenging for creating grammatically acceptable compressions. Finally, our automatic evaluation experiments motivate the use of relations-based F-score as a means of directly optimising compression quality, much in the same way MT systems optimise model parameters using BLEU as a measure of translation quality.

Acknowledgements

We are grateful to our annotators Vasilis Karaiskos, Beata Kouchnir, and Sarah Luger. Thanks to Jean Carletta, Frank Keller, Steve Renals, and Sebastian Riedel for helpful comments and suggestions. Lapata acknowledges the support of EPSRC (grant GR/T04540/01).

References

Bangalore, Srinivas, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st INLG*. Mitzpe Ramon, Israel, pages 1–8.

Briscoe, E. J. and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd LREC*. Las Palmas, Spain, pages 1499–1504.

Burnard, Lou. 2000. *The Users Reference Guide for the British National Corpus (World Edition)*. British National Corpus Consortium, Oxford University Computing Service.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st NAACL*. San Francisco, CA, pages 132–139.

Clarkson, Philip and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU–cambridge toolkit. In *Proceedings of Eurospeech*. Rhodes, Greece, pages 2707–2710.

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

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

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

Jing, Hongyan. 2000. Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th ANLP*. Seattle,WA, pages 310–315.

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

McDonald, Ryan. 2006. Discriminative sentence compression with soft syntactic constraints. In *Proceedings of the 11th EACL*. Trento, Italy, pages 297–304.

Nguyen, Minh Le, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho. 2004a. Example-based sentence reduction using the hidden Markov model. *ACM TALIP* 3(2):146–158.

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

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA.

Quinlan, J. R. 1993. *C4.5 – Programs for Machine Learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufman Publishers.

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

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

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

Weiss, Sholom M. and Casimir A. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.