Leveraging Structural Relations for Fluent Compressions at Multiple Compression Rates

Sourish Chaudhuri, Naman K. Gupta, Noah A. Smith, Carolyn P. Rosé
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA-15213, USA.
{sourishc, nkgupta, nasmith, cprose}@cs.cmu.edu

Abstract

Prior approaches to sentence compression have taken low level syntactic constraints into account in order to maintain grammaticality. We propose and successfully evaluate a more comprehensive, generalizable feature set that takes syntactic and structural relationships into account in order to sustain variable compression rates while making compressed sentences more coherent, grammatical and readable.

1 Introduction

We present an evaluation of the effect of syntactic and structural constraints at multiple levels of granularity on the robustness of sentence compression at varying compression rates. Our evaluation demonstrates that the new feature set produces significantly improved compressions across a range of compression rates compared to existing state-of-the-art approaches. Thus, we name our system for generating compressions the Adjustable Rate Compressor (ARC).

Knight and Marcu (2000) (K&M, henceforth) presented two approaches to the sentence compression problem: one using a noisy channel model, the other using a decision-based model. The performances of the two models were comparable though their experiments suggested that the noisy channel model degraded more smoothly than the decision-based model when tested on out-of-domain data. Riezler et al. (2003) applied linguistically rich LFG grammars to a sentence compression system. Turner and Charniak (2005) achieved similar performance to K&M using an unsupervised approach that induced rules from the Penn Treebank.

A variety of feature encodings have previously been explored for the problem of sentence compression. Clarke and Lapata (2007) included discourse level features in their framework to leverage context for enhancing coherence. McDonald's (2006) model (M06, henceforth) is similar to K&M except that it uses discriminative online learning to train feature weights. A key aspect of the M06 approach is a decoding algorithm that searches the entire space of compressions using dynamic programming to choose the best compression (details in Section 2). We use M06 as a foundation for this work because its soft constraint approach allows for natural integration of additional classes of features. Similar to most previous approaches, our approach compresses sentences by deleting words only.

The remainder of the paper is organized as follows. Section 2 discusses the architectural framework. Section 3 describes the innovations in the proposed model. We conclude after presenting the results of our evaluation in Section 4.

2 Experimental Paradigm

Supervised approaches to sentence compression typically use parallel corpora consisting of original and compressed sentences (paired corpus, henceforth). In this paper, we will refer to these pairs as a 2-tuple <*x*, *y*>, where *x* is the original sentence and *y* is the compressed sentence.

We implemented the M06 system as an experimental framework in which to conduct our investigation. The system uses as input the paired corpus, the corresponding POS tagged corpus, the paired corpus parsed using the Charniak parser (Charniak, 2000), and dependency parses from the MST parser (McDonald et al., 2005). Features are extracted over adjacent pairs of words in the compressed sentence and weights are learnt at training time using the MIRA algorithm (Crammer and Singer, 2003). We decode as follows to find the best compression:

Let the score of a compression *y* for a sentence *x* be *s*(*x*, *y*). This score is factored using a first-order Markov assumption over the words in the compressed sentence, and is defined by the dot product between a high dimensional feature representation and a corresponding weight vector (for details, refer to McDonald, 2006). The equations for decoding are as follows:

$$C[1] = 0.0$$
$$C[i] = \max_{j<i} C[j] + s(x, j, i), \forall i > 1$$

where C is the dynamic programming table and $C[i]$ represents the highest score for compressions ending at word i for the sentence x.

The M06 system takes the best scoring compression from the set of all possible compressions. In the ARC system, the model determines the compression rate and enforces a target compression length by altering the dynamic programming algorithm as suggested by M06:

$$C[1][1] = 0.0$$

$$C[1][r] = -\infty, \forall r > 1$$

$\forall i > 1,$

$$C[i][r] = \max_{j<i} C[j][r-1] + s(x, j, i)$$

where C is the dynamic programming table as before and $C[i][r]$ is the score for the best compression of length r that ends at position i in the sentence x. This algorithm runs in $O(n^2 r)$ time.

We define the rate of human generated compressions in the training corpus as the gold standard compression rate (GSCR). We train a linear regression model over the training data to predict the GSCR for a sentence based on the ratio between the lengths of each compressed-original sentence pair in the training set. The predicted compression rate is used to force the system to compress sentences in the test set to a specific target length. Based on the computed regression, the formula for computing the Predicted Compression Rate (PCR) from the Original Sentence Length (OSL) is as follows:

$$PCR = 0.86 - 0.004 \times OSL$$

In our work, enforcing specific compression rates serves two purposes. First, it allows us to make a more controlled comparison across approaches, since variation in compression rate across approaches confounds comparison of other aspects of performance. Second, it allows us to investigate how alternative models work at higher compression rates. Here our primary contribution is of robustness of the approach with respect to alternative feature spaces and compression rates.

3 Extended Feature Set

A major focus of our work is the inclusion of new types of features derived from syntactic analyses in order to make the resulting compressions more grammatical and thus increase the versatility of the resulting compression models.

The M06 system uses features extracted from the POS tagged paired corpus: POS bigrams, POS context of the words added to or dropped from the compression, and other information about the dropped words. For a more detailed description, please refer to McDonald, 2006.

From the phrase structure trees, M06 extracts context information about nodes that subsume dropped words. These features attempt to approximately encode changes in the grammar rules between source and target sentences. Dependency features include information about the dropped words' parents as well as conjunction features of the word and the parent.

Our extensions to the M06 feature set are inspired by an analysis of the compressions generated by it, and allow for a richer encoding of dropped words and phrases using properties of the words and their syntactic relations to the rest of the sentence. Consider this example (dropped words are *marked as such*):

** 68000 Sweden AB of Uppsala , Sweden , introduced* the *TeleServe , an integrated* answering machine and voice-message handler that links a Macintosh to Touch-Tone *phones* .

Note in the above example that the syntactic head of the sentence *introduced* has been dropped. Using the dependency parse, we add a class of features to be learned during training that lets the system decide when to drop the syntactic head of the sentence. Also note that *answering machine* in the original sentence was preceded by *an* while the word *the* was used with *Teleserve* (dropped in the compression). While POS information helps the system to learn that *the answering machine* is a good POS sequence, we do not have information that links the correct article to the noun. Information from the dependency parse allows us to learn when we can drop words whose heads are retained and when we can drop a head and still retain the dependent.

Now, consider the following example:

Examples for editors are applicable to awk patterns , grep and egrep .

Here, *Examples* has been dropped, while *for editors* which has *Examples* as a head is retained. Besides, in the sequence, *editors are applicable…*, the word *editors* behaves as the subject of *are* although the correct compression would have *examples* as its subject. A change in the arguments of the verbs will distort the meaning of the sentence. We augmented the feature set to include a class of features about structural information that tells us when the subject (or object) of a verb can be dropped while the verb itself is retained. Thus, now if the system does retain the *are*, it is more likely to retain the correct arguments of the word from the original sentence.

The new classes of features use only the dependency labels generated by the parser and are not lexicalized. Intuitively, these features help create units within the sentences that are tightly bound together, e.g., a subject and an object with its parent verb. We notice, as one would expect, that some dependency bindings are less strong than others. For instance, when faced with a choice, our system drops a relative pronoun thus breaking the dependency between the retained noun and the relative pronoun, rather than drop the noun, which was the retained subject.

Below is a summary of the information that the new features in our system encode:

[***Parent-Child***]- When a word is dropped, is its parent retained in the compression?

[***Dependent***]- When a word is dropped, are other words dependent on it (its children) also dropped or are they retained?

[***Verb-Arg***]- Information from the dependency parse about the subjects and objects of verbs can be used to encode more specific features (similar to the above) that say whether or not the subject (or object) was retained when the verb was dropped.

[***Sent-Head-Dep***]- Is the syntactic head of a sentence dropped?

4 Evaluation

We evaluate our model in comparison with M06. At training time, compression rates were not enforced on the ARC or M06 model. Our evaluation demonstrates that the proposed feature set produces more grammatical sentences across varying compression rates. In this section, GSCR denotes *gold standard compression rate* (i.e., the compression rate found in training data), CR denotes *compression rate*.

4.1 Corpora

Sentence compression systems have been tested on product review data from the Ziff-Davis (ZD, henceforth) Corpus by Knight and Marcu (2000), general news articles by Clarke and Lapata (CL, henceforth) corpus (2007) and biomedical articles (Lin and Wilbur, 2007). To evaluate our system, we used 2 test sets: **Set 1** contained 50 sentences; all 32 sentences from the ZD test set and 18 additional sentences chosen randomly from the CL test set; **Set 2** contained 40 sentences selected from the CL corpus, 20 of which were compressed at 75% of GSCR and 20 at 50% of GSCR (the percentages denote the enforced compression rates).

Three examples comparing compressed sentences are given below:

Original: *Like FaceLift, much of ATM 's screen performance depends on the underlying application.*

Human: *Much of ATM 's performance depends on the underlying application .*

M06: *'s screen performance depends on application*

ARC: *ATM 's screen performance depends on the underlying application .*

Original: *The discounted package for the Sparcserver 470 is priced at $89,900 , down from the regular $107,795 .*

Human: *The Sparcserver 470 is priced at $89,900 , down from the regular $107,795 .*

M06: *Sparcserver 470 is $89,900 regular $107,795*

ARC: *The discounted package is priced at $89,900 , regular $107,795 .*

The example below has compressions at 50% compression rate for M06 and ARC systems:

Original: *Cutbacks in local defence establishments is also a factor in some constituencies .*

M06: *establishments is a factor in some constituencies .*

ARC: *Cutbacks is a factor in some constituencies .*

Note that the subject of *is* is correctly retained in the ARC system.

4.2 User Study

In order to evaluate the effect of the features that we added to create the ARC model, we conducted a user study, adopting an experimental methodology similar to that used by K&M and M06. Each of four human judges, who were native speakers of English and not involved in the research we report in this paper, were instructed to rate two different sets of compressions along two dimensions, namely *Grammaticality* and *Completeness*, on a scale of 1 to 5. We chose to replace *Importance* (used by K&M), which is a task specific and possibly user specific notion, with the more general notion of *Completeness*, defined as the extent to which the compressed sentence is a complete sentence and communicates the main idea of the original sentence.

For Set 1, raters were given the original sentence and 4 compressed versions (presented in random order as in the M06 evaluation): the human compression, the compression produced by the original M06 system, the compression from the M06 system with GSCR, and the ARC system with GSCR. For Set 2, raters were given the original sentence, this time with two compressed versions, one from the M06 system and one from the ARC system, which were presented in a random order. Table 1 presents all the results in terms of human ratings of Grammaticality and Completeness as well as automatically computed ROUGE F_1 scores (Lin and Hovy, 2003). The scores in parentheses denote standard deviations.

	Grammaticality (Human Scores)	Completeness (Human Scores)	ROUGE F_1
Gold Standard	4.60 (0.69)	3.80(.99)	1.00 (0)
ARC (GSCR)	3.70 (1.10)	3.50(1.10)	.72 (.18)
M06	3.50 (1.30)	3.10(1.30)	.70 (.20)
M06 (GSCR)	3.10 (1.10)	3.10(1.10)	.71 (.18)
ARC (75%CR)	2.60 (1.10)	2.60(1.10)	.72 (.14)
M06 (75%CR)	2.20 (1.20)	2.00(1.00)	.67 (.20)
ARC (50%CR)	2.30 (1.30)	1.90(1.00)	.54 (.22)
M06 (50%CR)	1.90 (1.10)	1.80(1.00)	.58 (.22)

Table 1: Results of human judgments and ROUGE F_1

ROUGE scores were determined to have a significant positive correlation both with Grammaticality ($R = .46$, $p < .0001$) and Completeness ($R = .39$, $p < .0001$) when averaging across the 4 judges' ratings. On Set 1, a 2-tailed paired *t*-test reveals similar patterns for Grammaticality and Completeness: the human compressions are significantly better than any of the systems. ARC is significantly better than M06, both with enforced GSCR and without. M06 without GSCR is significantly better than M06 with GSCR. In Set 2 (with 75% and 50% GSCR enforced), the quality of compressions degrade as compression rate is made more severe; however, the ARC model consistently outperforms the M06 model with a statistically significant margin across compression rates on both evaluation criteria.

5 Conclusions and Future Work

In this paper, we designed a set of new classes of features to generate better compressions, and they were found to produce statistically significant improvements over the state-of-the-art. However, although the user study demonstrates the expected positive impact of grammatical features, an error analysis (Gupta et al., 2009) reveals some limitations to improvements that can be obtained using grammatical features that refer only to the source sentence structure, since the syntax of the source sentence is frequently not preserved in the gold standard compression. In our future work, we hope to explore alternative approaches that allow reordering or paraphrasing along with deleting words to make compressed sentences more grammatical and coherent.

Acknowledgments

The authors thank Kevin Knight and Daniel Marcu for sharing the Ziff-Davis corpus as well as the output of their systems, and the anonymous reviewers for their comments. This work was supported by the Cognitive and Neural Sciences Division, grant number N00014-00-1-0600.

References

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of NAACL*.

James Clarke and Mirella Lapata, 2007. Modelling Compression With Discourse Constraints. In *Proc. of EMNLP-CoNLL*.

Koby Crammer and Y. Singer. 2003. Ultraconservative online algorithms for multi-class problems. *JMLR*.

Naman K. Gupta, Sourish Chaudhuri and Carolyn P. Rosé, 2009. Evaluating the Syntactic Transformations in Gold Standard Corpora for Statistical Sentence Compression . In *Proc. of HLT-NAACL*.

Kevin Knight and Daniel Marcu. 2000. Statistics-Based Summarization – Step One: Sentence Compression. In *Proc. of AAAI*.

Jimmy Lin and W. John Wilbur. 2007. Syntactic sentence compression in the biomedical domain: facilitating access to related articles. *Information Retrieval*, 10(4):393-414.

Chin-Yew Lin and Eduard H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proc. of HLT-NAACL*.

Ryan McDonald, 2006. Discriminative sentence compression with soft syntactic constraints. In *Proc. of EACL*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proc.of ACL*.

S. Riezler, T. H. King, R. Crouch, and A. Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proc. of HLT-NAACL*.