Evaluating sentence compression: Pitfalls and suggested remedies

Courtney Napoles[1] and **Benjamin Van Durme**[1,2] and **Chris Callison-Burch**[1]
[1]Department of Computer Science
[2]Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

This work surveys existing evaluation methodologies for the task of sentence compression, identifies their shortcomings, and proposes alternatives. In particular, we examine the problems of evaluating paraphrastic compression and comparing the output of different models. We demonstrate that compression rate is a strong predictor of compression *quality* and that perceived improvement over other models is often a side effect of producing longer output.

1 Introduction

Sentence compression is the natural language generation (NLG) task of automatically shortening sentences. Because good compressions should be grammatical and retain important meaning, they must be evaluated along these two dimensions. Evaluation is a difficult problem for NLG, and many of the problems identified in this work are relevant for other generation tasks. Shared tasks are popular in many areas as a way to compare system performance in an unbiased manner. Unlike other tasks, such as machine translation, there is no shared-task evaluation for compression, even though some compression systems are indirectly evaluated as a part of DUC. The benefits of shared-task evaluation have been discussed before (e.g., Belz and Kilgarriff (2006) and Reiter and Belz (2006)), and they include comparing systems fairly under the same conditions.

One difficulty in evaluating compression systems fairly is that an unbiased automatic metric is hard to define. Automatic evaluation relies on a comparison to a single gold standard at a predetermined length, which greatly limits the types of compressions that can be fairly judged. As we will discuss in Section 2.1.1, automatic evaluation assumes that deletions are independent, considers only a single gold standard, and cannot handle compressions with paraphrasing. Like for most areas in NLG, human evaluation is preferable. However, as we discuss in Section 2.2, there are some subtleties to appropriate experiment design, which can give misleading results if not handled properly.

This work identifies the shortcomings of widely practiced evaluation methodologies and proposes alternatives. We report on the effect of compression rate on perceived quality and suggest ways to control for this dependency when evaluating across different systems. In this work we:

- highlight the importance of comparing systems with similar compression rates,
- argue that comparisons in many previous publications are invalid,
- provide suggestions for unbiased evaluation.

While many may find this discussion intuitive, these points are not addressed in much of the existing research, and therefore it is crucial to enumerate them in order to improve the scientific validity of the task.

2 Current Practices

Because it was developed in support of extractive summarization (Knight and Marcu, 2000), compression has mostly been framed as a deletion task (e.g., McDonald (2006), Galanis and Androutsopoulos (2010), Clarke and Lapata (2008), and Galley

Workshop on Monolingual Text-To-Text Generation, pages 91–97,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 91–97,
Portland, Oregon, 24 June 2011.

Words	Sentence
31	*Kaczynski faces charges contained in a 10-count federal indictment naming him as the person responsible for transporting bombs and bomb parts from Montana to California and mailing them to victims .*
17	Kaczynski faces charges naming him responsible for transporting bombs to California and mailing them to victims .
18	Kaczynski faces charges naming him responsible for transporting bombs and bomb parts and mailing them to victims .
18	Kaczynski faces a 10-count federal indictment for transporting bombs and bomb parts and mailing them to victims .

Table 1: Three acceptable compressions of a sentence created by different annotators (the first is the original).

and McKeown (2007)). In this context, a compression is an extracted subset of words from a long sentence. There are limited compression corpora because, even when an aligned corpus exists, the number of extractive sentence pairs will be few and therefore gold-standard compressions must be manually annotated. The most popular corpora are the Ziff-Davis corpus (Knight and Marcu, 2000), which contains a small set of 1067 extracted sentences from article/abstract pairs, and the manually annotated Clarke and Lapata (2008) corpus, consisting of nearly 3000 sentences from news articles and broadcast news transcripts. These corpora contain one gold standard for each sentence.

2.1 Automatic Techniques

One of the most widely used automatic metrics is the F1 measure over grammatical relations of the gold-standard compressions (Riezler et al., 2003). This metric correlates significantly with human judgments and is better than Simple String Accuracy (Bangalore et al., 2000) for judging compression quality (Clarke and Lapata, 2006). F1 has also been used over unigrams (Martins and Smith, 2009) and bigrams (Unno et al., 2006). Unno et al. (2006) compared the F1 measures to BLEU scores (using the gold standard as a single reference) over varying compression rates, and found that BLEU behaves similarly to both F1 measures. A syntactic approach considers the alignment over parse trees (Jing, 2000), and a similar technique has been used with dependency trees to evaluate the quality of sentence fusions (Marsi and Krahmer, 2005).

The only metric that has been shown to correlate with human judgments is F1 (Clarke and Lapata, 2006), but even this is not entirely reliable. F1 over grammatical relations also depends on parser accuracy and the type of dependency relations used.[1]

2.1.1 Pitfalls of Automatic Evaluation

Automatic evaluation operates under three often incorrect assumptions:

Deletions are independent. The dependency structure of a sentence may be unaltered when dependent words are not deleted as a unit. Examples of words that should be treated as a single unit include negations and negative polarity items or certain multi-word phrases (such as deleting *Latin* and leaving *America*). F1 treats all deletions equally, when in fact errors of this type may dramatically alter the meaning or the grammaticality of a sentence and should be penalized more than less serious errors, such as deleting an article.

The gold standard is the single best compression. Automatic evaluation considers a single gold-standard compression. This ignores the possibility of different length compressions and equally good compressions of the same length, where multiple non-overlapping deletions are acceptable. For an example, see Table 1.

Having multiple gold standards would provide references at different compression lengths and reflect different deletion choices (see Section 3). Since no large corpus with multiple gold standards exists to our knowledge, systems could instead report the quality of compressions at several different compression rates, as Nomoto (2008) did. Alternatively, systems could evaluate compressions that are of a similar length as the gold standard compression, to fix a length for the purpose of evaluation. Output length is controlled for evaluation in some other areas, notably DUC.

Systems compress by deletion and not substitution. More recent approaches to compression introduce reordering and paraphrase operations (e.g.,

[1]For example, the RASP parser uses 16 grammatical dependencies (Briscoe, 2006) while there are over 50 Stanford Dependencies (de Marneffe and Manning, 2008).

Cohn and Lapata (2008), Woodsend et al. (2010), and Napoles et al. (2011)). For paraphrastic compressions, manual evaluation alone reliably determines the compression quality. Because automatic evaluation metrics compare shortened sentences to extractive gold standards, they cannot be applied to paraphrastic compression.

To apply automatic techniques to substitution-based compression, one would need a gold-standard set of paraphrastic compressions. These are rare. Cohn and Lapata (2008) created an abstractive corpus, which contains word reordering and paraphrasing in addition to deletion. Unfortunately, this corpus is small (575 sentences) and only includes one possible compression for each sentence.

Other alternatives include deriving such corpora from existing corpora of multi-reference translations. The longest reference translation can be paired with the shortest reference to represent a long sentence and corresponding paraphrased gold-standard compression.

Similar to machine translation or summarization, automatic translation of paraphrastic compressions would require *multiple references* to capture allowable variation, since there are often many equally valid ways of compressing an input. ROUGE or BLEU could be applied to a set of multiple-reference compressions, although BLEU is not without its own shortcomings (Callison-Burch et al., 2006). One benefit of both ROUGE and BLEU is that they are based on n-gram recall and precision (respectively) instead of word-error rate, so reordering and word substitutions can be evaluated. Dorr et al. (2003) used BLEU for evaluation in the context of headline generation, which uses rewording and is related to sentence compression. Alternatively, manual evalation can be adapted from other NLG domains, such as the techniques described in the following section.

2.2 Manual Evaluation

In order to determine semantic and syntactic suitability, manual evaluation is preferable over automatic techniques whenever possible. The most widely practiced manual evaluation methodology was first used by Knight and Marcu (2002). Judges grade each compressed sentence against the original and make two separate decisions: how grammatical is the compression and how much of the meaning from the original sentence is preserved. Decisions are rated along a 5-point scale (LDC, 2005).

Most compression systems consider sentences out of context (a few exceptions exist, e.g., Daumé III and Marcu (2002), Martins and Smith (2009), and Lin (2003)). Contextual cues and discourse structure may not be a factor to consider if the sentences are generated for use out of context. An example of a context-aware approach considered the summaries formed by shortened sentences and evaluated the compression systems based on how well people could answer questions about the original document from the summaries (Clarke and Lapata, 2007). This technique has been used before for evaluating summarization and text comprehension (Mani et al., 2002; Morris et al., 1992).

2.2.1 Pitfalls of Manual Evaluation

Grammar judgments decrease when the compression is presented alongside the original sentence. Figure 1 shows that the mean grammar rating for the same compressions is on average about 0.3 points higher when the compression is judged in isolation. Researchers should be careful to state when grammar is judged on compressions lacking reference sentences.

Another factor is the group of judges. Obviously different studies will rely on different judges, so whenever possible the sentences from an existing model should be re-evaluated alongside the new model. The "McD" entries in Table 2 represent a set of sentences generated from the exact same model evaluated by two different sets of judges. The mean grammar and meaning ratings in each evaluation setup differ by 0.5–0.7 points.

3 Compression Rate Predicts Performance

The dominant assumption in compression research is that the system makes the determination about the optimal compression length. For this reason, compression rates can vary drastically across systems. In order to get unbiased evaluations, systems should be compared only when they are compressing at similar rates.

Compression rate is defined as:

$$\frac{\text{\# of tokens in compressed sentence}}{\text{\# of tokens in original sentence}} \times 100 \qquad (1)$$

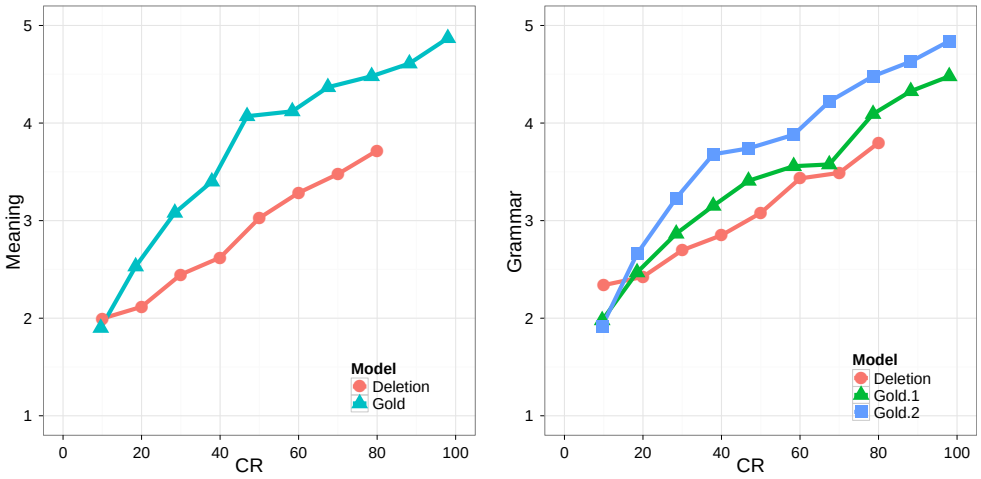

Figure 1: Compression rate strongly correlates with human judgments of meaning and grammaticality. *Gold* represents gold-standard compression and *Deletion* the results of a leading deletion model. Gold.1 grammar judgments were made alongside the original sentence and Gold.2 were made in isolation.

It seems intuitive that sentence quality diminishes in relation to the compression rate. Each word deleted increases the probability that errors are introduced. To verify this notion, we generated compressions at decreasing compression rates of 250 sentences randomly chosen from the written corpus of Clarke and Lapata (2008), generated by our implementation of a leading extractive compression system (Clarke and Lapata, 2008). We collected human judgments using the 5-point scales of meaning and grammar described above. Both quality judgments decreased linearly with the compression rate (see "Deletion" in Figure 1).

As this behavior could have been an artifact of the particular model employed, we next developed a unique gold-standard corpus for 50 sentences selected at random from the same corpus described above. The authors manually compressed each sentence at compression rates ranging from less than 10 to 100. Using the same setup as before, we collected human judgments of these gold standards to determine an upper bound of perceived quality at a wide range of compression rates. Figure 1 demonstrates that meaning and grammar ratings decay more drastically at compression rates below 40 (see "Gold"). Analysis suggests that humans are often able to practice "creative deletion" to tighten a sentence up to a certain point, before hitting a compression barrier, shortening beyond which leads to significant meaning and grammatically loss.

4 Mismatched Comparisons

We have observed that a difference in compression rates as small as 5 percentage points can influence the quality ratings by as much as 0.1 points and conclude: systems must be compared using similar levels of compression. In particular, if system A's output is higher quality, but longer than system B's, then it is not necessarily the case that A is better than B. Conversely, if B has results at least as good as system A, one can claim that B is better, since B's output is shorter.

Here are some examples in the literature of mismatched comparisons:

- Nomoto (2009) concluded their system significantly outperformed that of Cohn and Lapata (2008). However, the compression rate of their system ranged from 45 to 74, while the compression rate of Cohn and Lapata (2008) was 35. This claim is unverifiable without further comparison.
- Clarke and Lapata (2007), when comparing against McDonald (2006), reported significantly better results at a 5-point higher compression rate. At first glance, this does not seem like a remarkable difference. However,

Model	Meaning	Grammar	CompR
C&L	3.83	3.66	64.1
McD	3.94	3.87	64.2
C&L	3.76*	3.53*	78.4*
McD	3.50*	3.17*	68.5*

Table 2: Mean quality ratings of two competing models once the compression rates have been standardized, and as reported in the original work (denoted *). There is no significant improvement, but the numerically better model changes.

> the study evaluated the quality of summaries containing automatically shortened sentences. The average document length in the test set was 20 sentences, and with approximately 24 words per sentence, a typical 65.4% compressed document would have 80 more words than a typical 60.1% McDonald compression. The aggregate loss from 80 words can be considerable, which suggests that this comparison is inconclusive.

We re-evaluated the model described in Clarke and Lapata (2008) (henceforth C&L) against the McDonald (2006) model with global constraints, but fixed the compression rates to be equal. We randomly selected 100 sentences from that same corpus and generated compressions with the same compression rate as the sentences generated by the McDonald model (McD), using our implementation of C&L. Although not statistically significant, this new evaluation reversed the polarity of the results reported by Clarke and Lapata (Table 2). This again stresses the importance of using similar compression rates to draw accurate conclusions about different models.

An example of unbiased evaluation is found in Cohn and Lapata (2009). In this work, their model achieved results significantly better than a competing system (McDonald, 2006). Recognizing that their compression rate was about 15 percentage points higher than the competing system, they fixed the target compression rate to one similar to McDonald's output, and still found significantly better performance using automatic measures. This work is one of the few that controls their output length in order to make an objective comparison (another example is found in McDonald (2006)), and this type of analysis should be emulated in the future.

5 Suggestions

Models should be tested on the same corpus, because different corpora will likely have different features that make them easier or harder to compress. In order to make non-vacuous comparisons of different models, a system also needs to be constrained to produce the same length output as another system, or report results *at least as good* for shorter compressions. Using the multi-reference gold-standard collection described in Section 3, relative performance could be estimated through comparison to the gold-standard curve. The reference set we have annotated is yet small, but this is an area for future work based on feedback from the community.[2]

Other methods for limiting quality disparities introduced by the compression rate include fixing the target length to that of the gold standard (e.g., Unno et al. (2006)). Alternately, results for a system at varying compression levels can be reported,[3] allowing for comparisons at similar lengths. This is a practice to be emulated, if possible, because systems that cannot control output length can make comparisons against the appropriate compression rate.

In conclusion, we have provided justification for the following practices in evaluating compressions:

- Compare systems at similar compression rates.
- Provide results across multiple compression rates when possible.
- Report that system A surpasses B iff: A and B have the same compression rate and A does better than B, or A produces shorter output than B and A does at least as well B.
- New corpora for compression should have multiple gold standards for each sentence.

Acknowledgments

We are very grateful to James Clarke for helping us obtain the results of existing systems and to the reviewers for their helpful comments and recommendations. The first author was supported by the JHU Human Language Technology Center of Excellence. This research was funded in part by the NSF under grant IIS-0713448. The views and findings are the authors' alone.

[2]This data is available on request.

[3]For example, Nomoto (2008) reported results ranging over compression rates: 0.50–0.70.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 1–8. Association for Computational Linguistics.

A. Belz and A. Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 133–135. Association for Computational Linguistics.

Ted Briscoe. 2006. An introduction to tag sequence grammars and the RASP system parser. *Computer Laboratory Technical Report*, 662.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of EACL*, Trento, Italy.

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

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1–11.

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

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of COLING*.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

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

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL Workshop on Text summarization Workshop*.

Dimitrios Galanis and Ion Androutsopoulos. 2010. An extractive supervised two-stage method for sentence compression. In *Proceedings of NAACL*.

Michel Galley and Kathleen R. McKeown. 2007. Lexicalized Markov grammars for sentence compression. *the Proceedings of NAACL/HLT*.

Shudong Huang, David Graff, and George Doddington. 2002. Multiple-Translation Chinese Corpus. Linguistic Data Consortium.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pages 310–315. Association for Computational Linguistics.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization – Step one: Sentence compression. In *Proceedings of AAAI*.

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

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Chin-Yew Lin. 2003. Improving summarization performance by sentence compression: a pilot study. In *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, pages 1–8. Association for Computational Linguistics.

Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(01):43–68.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. In *Proceedings of the European Workshop on Natural Language Generation*, pages 8–10.

André F. T. Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*.

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

Andrew H. Morris, George M. Kasper, and Dennis A. Adams. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *INFORMATION SYSTEMS RESEARCH*, 3(1):17–35.

Courtney Napoles, Chris Callison-Burch, Juri Ganitkevitch, and Benjamin Van Durme. 2011. Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of ACL, Workshop on Monolingual Text-To-Text Generation*.

Tadashi Nomoto. 2008. A generic sentence trimmer with CRFs. *Proceedings of ACL-08: HLT*, pages 299–307.

Tadashi Nomoto. 2009. A comparison of model free versus model intensive approaches to sentence compression. In *Proceedings of EMNLP*.

E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 136–138. Association for Computational Linguistics.

Stefan Riezler, 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 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 118–125. Association for Computational Linguistics.

Yuya Unno, Takashi Ninomiya, Yusuke Miyao, and Jun'ichi Tsujii. 2006. Trimming CFG parse trees for sentence compression using machine learning approaches. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 850–857. Association for Computational Linguistics.

Kristian Woodsend, Yansong Feng, and Mirella Lapata. 2010. Generation with quasi-synchronous grammar. In *Proceedings of EMNLP*.