A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs

Kristina Toutanova
Microsoft Research
Redmond, WA, USA

Chris Brockett
Microsoft Research
Redmond, WA, USA

Ke M. Tran*
University of Amsterdam
Amsterdam, The Netherlands

Saleema Amershi
Microsoft Research
Redmond, WA, USA

Abstract

We introduce a manually-created, multi-reference dataset for abstractive sentence and short paragraph compression. First, we examine the impact of single- and multi-sentence level editing operations on human compression quality as found in this corpus. We observe that substitution and rephrasing operations are more meaning preserving than other operations, and that compressing in context improves quality. Second, we systematically explore the correlations between automatic evaluation metrics and human judgments of meaning preservation and grammaticality in the compression task, and analyze the impact of the linguistic units used and precision versus recall measures on the quality of the metrics. Multi-reference evaluation metrics are shown to offer significant advantage over single reference-based metrics.

1 Introduction

Automated sentence compression condenses a sentence or paragraph to its most important content in order to enhance writing quality, meet document length constraints, and build more accurate document summarization systems (Berg-Kirkpatrick et al., 2011; Vanderwende et al., 2007). Though word deletion is extensively used (e.g., (Clarke and Lapata, 2008)), state-of-the-art compression models (Cohn and Lapata, 2008; Rush et al., 2015) benefit crucially from data that can represent complex abstractive compression operations, including substitution of words and phrases and reordering.

This paper has two parts. In the first half, we introduce a manually-created *multi-reference* dataset for *abstractive* compression of sentences and short paragraphs, with the following features:

- It contains approximately 6,000 source texts with multiple compressions (about 26,000 pairs of source and compressed texts), representing business letters, newswire, journals, and technical documents sampled from the Open American National Corpus (OANC[1]).
- Each source text is accompanied by up to five crowd-sourced rewrites constrained to a preset compression ratio and annotated with quality judgments. Multiple rewrites permit study of the impact of operations on human compression quality and facilitate automatic evaluation.
- This dataset is the first to provide compressions at the multi-sentence (two-sentence paragraph) level, which may present a stepping stone to whole document summarization. Many of these two-sentence paragraphs are compressed both as paragraphs and separately sentence-by-sentence, offering data that may yield insights into the impact of multi-sentence operations on human compression quality.
- A detailed edit history is provided that may allow fine-grained alignment of original and compressed texts and measurement of the cognitive load of different rewrite operations.

Our analysis of this dataset reveals that abstraction has a significant positive impact on meaning preservation, and that application of trans-sentential

*This research was conducted during the author's internship at Microsoft Research.

[1] http://www.anc.org/data/oanc

context has a significant positive impact on both meaning preservation and grammaticality.

In the second part, we provide a systematic empirical study of eighty automatic evaluation metrics for text compression using this dataset, correlating them with human judgments of meaning and grammar. Our study shows strong correlation of the best metrics with human judgments of meaning, but weaker correlations with judgments of grammar. We demonstrate significant gains from multiple references. We also provide analyses of the impact of the linguistics units used (surface n-grams of different sizes versus parse-based triples), and the use of precision versus recall-based measures.

2 Related Work

Prior studies of human compression: Clarke (2008) studied the properties of manually-collected deletion-based compressions in the news genre, comparing them with automatically-mined data from the Ziff-Davis corpus in terms of compression rate, length of deleted spans, and deletion probability by syntactic constituent type. Jing and McKeown (1999) identified abstractive operations (other than word deletion) employed by professional writers, including paraphrasing and re-ordering of phrases, and merging and reordering sentences, but did not quantify their impact on compression quality.

Deletion-based compression corpora: Currently available automatically-mined deletion corpora are single-reference and have varying (uncontrolled) compression rates. Knight and Marcu (2002) automatically mined a small parallel corpus (1,035 training and 32 test sentences) by aligning abstracts to sentences in articles. Filippova and Altun (2013) extracted deletion-based compressions by aligning news headlines to first sentences, yielding a corpus of 250,000 parallel sentences. The same approach was used by Filippova et al. (2015) to create a set of 2M sentence pairs. Only a subset of 10,000 parallel sentences from the latter has been publicly released. Clarke and Lapata (2006) and Clarke and Lapata (2008) provide two manually-created two-reference corpora for deletion-based compression:[2] their sizes are 1,370 and 1,433 sentences, respectively.

[2] http://jamesclarke.net/research/resources

Abstractive compression corpora: Rush et al. (2015) have mined 4 million compression pairs from news articles and released their code to extract data from the Annotated Gigaword (Napoles et al., 2012). A news-domain parallel sentence corpus containing 1,496 parallel examples has been culled from multi-reference Chinese-English translations by Ganitkevitch et al. (2011). The only publicly-available manually-created abstractive compression corpus is that described by Cohn and Lapata (2008), which comprises 575 single-reference sentence pairs.

Automatic metrics: Early automatic metrics for evaluation of compressions include success rate (Jing, 2000), defined as accuracy of individual word or constituent deletion decisions; Simple String Accuracy (string edit distance), introduced by Bangalore et al. (2000) for natural language generation tasks; and Word Accuracy (Chiori and Furui, 2004), which generalizes Bangalore et al. (2000) to multiple references. Riezler et al. (2003) introduced the use of F-measure over grammatical relations. Word unigram and word-bigram F-measure have also been used (Unno et al., 2006; Filippova et al., 2015). Variants of ROUGE (Lin, 2004), used for summarization evaluation, have also been applied to sentence compressions (Rush et al., 2015).

Riezler et al. (2003) show that F-measure over grammatical relations agrees with human ratings on the relative ranking of three systems at the corpus level. Clarke and Lapata (2006) evaluate two deletion-based automatic compression systems against a deletion-based gold-standard on sets of 20 sentences. Parse-based F-1 was shown to have high sentence-level Pearson's ρ correlation with human judgments of overall quality, and to have higher ρ than Simple String Accuracy.

Napoles et al. (2011) have pointed to the need of multiple references and studies of evaluation metrics. For the related tasks of document and multi-document summarization, Graham (2015) provides a fine-grained comparison of automated evaluation methods. However, to the best of our knowledge, no studies of automatic evaluation metrics exist for abstractive compression of shorter texts.

Length		Text	Operations
1-Sent	Source	Think of all the ways everyone in your household will benefit from your membership in Audubon.	N/A
	Ref-1	Imagine how your household will benefit from your Audubon membership.	paraphrase + deletion + transformation
	Ref-2	Everyone in your household will benefit from membership in Audubon.	deletion
2-Sent	Source	Will the administration live up to its environmental promises? Can we save the last of our ancient forests from the chainsaw?	N/A
	Ref-1	Can the administration keep its promises? Can we save the last of our forests from loss?	two-sentences + deletion + paraphrase
	Ref-2	Will the administration live up to its environmental promises to save our ancient forests?	merge + deletion

Table 1: Examples of 1- and 2-sentence crowd-sourced compressions, illustrating different rewrite types.

	Newswire	Letters	Journal	Non-fiction
#texts	695	1,591	1,871	2,012

Table 2: Overview of the dataset by genre.

3 Dataset: Annotation and Properties

We sampled single sentences and two-sentence paragraphs from several genres in the written text section of the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) of the Open American National Corpus (OANC), supplemented by additional data from the written section of OANC. Two-sentence paragraphs account for approximately 23% of multi-sentence paragraphs in the OANC. The two-sentence paragraphs we sampled contain at least 25 words. Table 2 breaks the sampled texts down by genre. Non-news genres are better represented in our sample than the newswire typically used in compression tasks. The *Letters* examples are expected to be useful for learning to compress emails. The *Journal* texts are likely to be challenging as their purpose is often more than to convey information. The *Non-Fiction* collection includes material from technical academic publications, such as PLoS Medicine, an open access journal.[3]

3.1 Annotation

Compressions were created using UHRS, an in-house crowd-sourcing system similar to Amazon's Mechanical Turk, in two annotation rounds, one for shortening and a second to rate compression quality.

Generating compressions: In the first round, we asked five workers (*editors*) to abridge each source text by at least 25%, while remaining grammatical and fluent, and retaining the meaning of the original. This requirement was enforced programmatically on the basis of character count. The 25% rate is intended to reflect practical editing scenarios (e.g., shrink 8 pages to 6). To facilitate meeting this requirement, the minimum source text length presented to editors was 15 words. For a subset of paragraphs, we collected compressions both as independent rewrites of their component sentences, and of the paragraph as a whole. Table 1 show compression examples and strategies.

Evaluating compression quality: In the second round, we asked 3-5 judges (*raters*) to evaluate the grammaticality of each compression on a scale from 1 (major errors, disfluent) through 3 (fluent), and again analogously for meaning preservation on a scale from 1 (orthogonal) through 3 (most important meaning-preserving).[4] We later used the same process to evaluate compressions produced by automatic systems. The full guidelines for the editors and raters are available with the data release.

Quality controls: All editors and raters were based in the US, and the raters were required to pass a qualification test which asked them to rate the meaning and grammaticality for a set of examples with known answers. To further improve the quality of the data, we removed low-quality compressions. We computed the quality of each compression as the average of the grammar and meaning quality as judged by the raters. We then computed the mean quality for each editor, and removed compressions authored by the bottom 10% of editors. We did the same for the bottom 10% of the raters.[5]

[3] http://journals.plos.org/plosmedicine/

[4] Pilot studies suggested that a scale of 1-3 offered better inter-annotator agreement than the standard 5-point Likert-type scale, at the cost of granularity.

[5] This was motivated by the observation that the quality of work produced by judges is relatively constant (Gao et al., 2015).

Description	Texts			Quality	
	Source	Target	Avg CPS	Meaning	Grammar
All	6,169	26,423	4.28	2.78	2.82
Per Source Length					
1-Sent	3,764	15,523	4.12	2.78	2.81
2-Sent	2,405	10,900	4.53	2.78	2.83

Table 3: Overview of the dataset, presenting the overall number of source and target texts, the average quality of the compressed texts, and breakdown by length of source (number of sentences).

Table 3 shows the number of compressions in the cleaned dataset, as well as the average number of compressions per source text (CPS) and the average meaning and grammar scores. Meaning quality and grammaticality scores are relatively good, averaging 2.78 and 2.82 respectively. The filtered crowd-sourced compressions were most frequently judged to retain the most important meaning (80% of the time), or much of the meaning (17% of the time), with the lowest rating of 1 appearing only 3% of the time. This distribution is quite different from that of automatic compression systems in Section 4.

We provide a standard split of the data into training, development and test sets.[6] There are 4,936 source texts in the training, 448 in the development, and 785 in the test set.

3.2 Inter-Annotator Agreement

Crowd Workers: Since a different set of judges performs each task, large sets of inputs judged by the same two raters are unavailable. To simulate two raters, we follow Pavlick and Tetrault (2016): for each sentence, we randomly choose one annotator's output as the category for annotator A, and select the rounded average ranking for the remaining annotators as the category for annotator B. We then compute quadratic weighted κ (Cohen, 1968) for this pair over the whole corpus. We repeat the process 1000 times to compute the mean and variance of κ. The first row of the Table 4 reports the absolute agreement and κ, where the absolute agreement measures the fraction of times that A is equal to B. The 95% confidence intervals for κ are narrow, with width at most .01.

[6]The dataset can be downloaded from the project's website `https://www.microsoft.com/en-us/research/project/intelligent-editing/`.

Description	**Meaning**		**Grammar**	
	Agreement	κ	**Agreement**	κ
worker versus worker	.721	.306	.784	.381
expert versus expert	.888	.518	.890	.514
expert versus worker	.946	.549	.930	.344

Table 4: Agreement on meaning preservation and grammatically between crowd workers and experts.

Expert Raters: A small sample of 116 sentence pairs was rated by two expert judges. We used quadratic weighted κ directly, without sampling. To assess agreement between experts and non-experts, we computed weighted κ between the (rounded) average of the expert judgments and the (rounded) average of the crowd judgments, using 25,000 bootstrap replications each. The results are shown in the last two rows of Table 4. The confidence intervals for κ are wide due to the small sample size, and span values up to .17 away from the mean. Overall, agreement of experts with the average crowd-sourced ratings is moderate (approaching substantial) for meaning, and fair for grammar.

3.3 Analysis of Editing Operations

Frequency analysis: To analyze the editing operations used, we applied the state-of-the-art monolingual aligner Jacana (Yao et al., 2013) to align input to compressed texts. Out of the 26,423 compressions collected, 25.2% contained only token deletions. Those containing deletion and reordering amounted to a mere 9.1%, while those that also contain substitution or rephrasing (abstractive compressions) is 65.6%. Although abstraction is present in the large majority of compressions, these statistics do not indicate that paraphrasing is more prevalent than copying at the token level. The word alignments for target compression words indicate that 7.1% of target tokens were inserted, 75.4% were copied and 17.3% were paraphrased. From the alignments for source text words, we see that 31% of source words were deleted. The fraction of inserted and deleted words is probably overestimated by this approach, as it is likely that sequences of source words were abstracted as shorter sequences of target words in many-to-one or many-to-many alignment patterns that are difficult to detect automatically.

For the subset of examples where the input text

Operation	Meaning		Grammar	
	Present	Absent	Present	Absent
Substitute	2.81**	2.70	2.79	2.85**
Reorder	2.80	2.82	2.80	2.82**
Merge	2.63	2.82**	2.84**	2.82
Sentence Delete	2.57	2.82*	2.84	2.75

Table 5: Meaning and grammaticality judgments by compression operation. *p = 0.002. **p $<$ 0.0001.

Source Type	Meaning	Grammar
2-Sentence	2.86**	2.87**
1-Sentence	2.78	2.82

Table 6: Meaning and grammaticality judgments for compressing two sentences jointly versus individually. **p $<$ 0.0001.

contained more than one sentence, we computed the frequency of sentence-merging and sentence deletion when compressing. Of the compressions for two-sentence paragraphs, 72.4% had two sentences in the output, 0.4% had one sentence deleted, and 27.3% had the two source sentences merged.

Impact of operations: Because the dataset contains multiple compressions of the same sources, we are able to estimate the impact of different editing operations. These were classified using the Jacana word alignment tool. Table 5 presents the average judgment scores for meaning preservation and grammaticality for four operations. The upper two rows apply to all texts, the lower two to two-sentence paragraphs only. The statistical significance of their impact was tested using the Wilcoxon signed-rank test on paired observations. It appears that raters view compressions that involve substitutions as significantly more meaning-preserving than those that do not ($p < 0.0001$), but judge their grammaticality to be lower than that of deletion-based compressions. Note that the reduced grammaticality may be due to typographical errors that have been introduced during rephrasing, which could have been avoided had a more powerful word processor been used as an editing platform. Reordering has no significant impact on meaning, but leads to substantial degradation in grammatically. Conversely, abridgments that merge or delete sentences are rated as significantly less meaning preserving, but score higher for grammaticality, possibly reflecting greater skill on the part of those editors..

Impact of sentence context: Table 6 shows that the context provided by 2-sentence sources yields significantly improved scores for both meaning and grammaticality. Here we used the matched pairs design to compare the average quality of two-sentence paragraph compressions with the average quality of the compressions of the same paragraphs produced by separately compressing the two sentences.

4 Evaluating Evaluation Metrics

Progress in automated text compression is standardly measured by comparing model outputs at the *corpus* level. To train models discriminatively and to perform fine-grained system comparisons, however, it is also necessary to have evaluation of system outputs at the *individual input* level. Below, we examine automated metric correlation with human judgments at both levels of granularity.

4.1 Automatic Metrics

The goal of this analysis is to develop an understanding of the performance of automatic evaluation metrics for text compression, and the factors contributing to their performance. To this end, we group automatic metrics according to three criteria. The first is the *linguistic units* used to compare system and reference compressions. Prior work on compression evaluation has indicated that a parse-based metric is superior to one based on surface substrings (Clarke and Lapata, 2006), but the contribution of the linguistic units has not been isolated, and surface n-gram units have otherwise been successfully used for evaluation in related tasks (Graham, 2015). Accordingly, we empirically compare metrics based on surface uni-grams (LR-1), bi-grams (LR-2), tri-grams (LR-3), and four-grams (LR-4), as well skip bi-grams (with a maximum of four intervening words as in ROUGE-S4) (SKIP-2), and dependency tree triples obtained from collapsed dependencies output from the Stanford parser (PARSE-2).[7] The second criterion is the *scoring measure* used to evaluate the match between two sets of linguistic units corresponding to a system output and a reference compression. We compare Precision, Recall, F-measure, and Precision+Brevity penalty (as

[7] Clarke and Lapata (2006) used the RASP parser (Briscoe and Carroll, 2002), but we expect that the Stanford parser is similarly robust and would lead to similar correlations.

in BLEU). The third criterion is whether *multiple references* or a *single reference* is used, and in the case of multiple references, the method used to aggregate information from multiple references. We investigate two previously applied methods and introduce a novel approach that often outperforms the standard methods.

To illustrate, we introduce some notation and use a simple example. Consider a sub-phrase of one of the sentences in Table 1, *think about your household*, as an input text to compress. Let us assume that we have two reference compressions, R1: *imagine your household*, and R2: *your household*. Each metric m is a function from a pair of a system output o and a list of references $r_1, r_2, \ldots, r_k$ to the reals. To compute most metrics, we first compute a linguistic unit feature vector for each reference $\Phi(r_j)$, as well as for the set of references $\Phi(r_1, r_2, \ldots, r_k)$. Similarly, we compute a linguistic unit vector for the output $\Phi(o)$ and measure the overlap between the system and reference vectors. The vectors of the example references, if we use surface bigram units, would be, for R1, {`imagine_your:1`, `your_household:1`}, and for R2, {`your_household:1`}. The weights of all n-grams in individual references and system outputs are equal to 1.[8] If we use dependency-parse triples instead, the vector of R2 would be {`nmod:poss(household, your):1`}.

The precision of a system output against a reference is defined as the match $\Phi(r)^T\Phi(o)$ divided by the number of units in the vector of o; the latter can be expressed as the L_1 norm of $\Phi(o)$ because all weights are positive: $\text{Precision}(o, r) = \frac{\Phi(r)^T\Phi(o)}{|\Phi(o)|_1}$. The recall against a single reference can be similarly defined as the match divided by the number of units in the reference: $\text{Recall}(o, r) = \frac{\Phi(r)^T\Phi(o)}{|\Phi(r)|_1}$.

We distinguish three methods for aggregating information from multiple references: MULT-MAX which uses the single reference out of a set that results in the highest single-reference score, and two further methods, MULT-ALL and MULT-PROB, that construct an aggregate linguistic unit vector $\Phi(r_1, \ldots, r_k)$ before matching. MULT-ALL is the standard method used in multi-reference BLEU, where the vector for a set of references is defined as the union of the features of the set. For our example, the combined vector of R1 and R2 is equal to the vector of R1, because R2 adds no new bigrams. MULT-PROB, a new method that we propose here, is motivated by the observation that although judgments of importance of content are subjective, the more annotators assert some information is important, the more this information should contribute to the matching score.[9] In MULT-PROB we define the weight of a linguistic unit in the combined reference vector as the proportion of references that include the unit. For our example, $\Phi_{\text{MULT-PROB}}(R1, R2)$ is {`imagine_your`:.5, `your_household`:1}.

4.2 Models for Text Compression

For the purpose of analysis, we trained and evaluated four compression systems. These include both deletion-based and abstractive models: (1) ILP, an integer linear programing approach for deletion-based compression (Clarke and Lapata, 2008), (2) T3, a tree transducer-based model for abstractive compression (Cohn and Lapata, 2008), (3) Seq2seq, a neural network model for deletion-based compression (Filippova et al., 2015), and (4) NAMAS, a neural model for abstractive compression and summarization (Rush et al., 2015). We are not concerned with the relative performance of these models so much as we are concerned with evaluating the automatic evaluation metrics themselves. We have sought to make the models competitive, but have not required that all systems use identical training data.

All of the models are evaluated on the test set portion of our dataset. All models use the training portion of the data for training, and two models (Seq2Seq and NAMAS[10]) additionally use external training data. The external data is summarized in Table 7. The Gigaword set was extracted from the Annotated Gigaword (Napoles et al., 2012), using the implementation provided by Rush et al. (2015). The Headline data was extracted in similar fashion using an in-house news collection.

[8] We handle repeating n-grams by assigning each subsequent n-gram of the same type a distinct type, so that the i-th *the* of a system output can match the i-th *the* of a reference.

[9] A similar insight was used in one of the component metrics of the SARI evaluation metric used for text simplification evaluation (Xu et al., 2016).

[10] The original works introducing these models employed much larger training corpora, believed to be key to improving the accuracy of neutral network models with large parameter spaces.

Data		#src tokens	#trg tokens	#sents
Abstractive	Gigaword	114.1M	30.0M	3.6M
	Headline	6.0M	1.4M	0.2M
Deletion-based	Gigaword	1,353K	329K	47K
	Headline	59K	11K	2K

Table 7: External data statistics.

ILP: We use an open-source implementation[11] of the semi-supervised ILP model described in (Clarke and Lapata, 2008). The model uses a trigram language model trained on a 9 million token subset of the OANC corpus. The ILP model requires parsed sentences coupled with deletion-based compressions for training, so we filtered and preprocessed our dataset to satisfy these constraints. We used all single sentence inputs with their corresponding deletion-based compressions, and additionally used two-sentence paragraph input/output pairs split into sentences by heuristically aligning source to target sentences in the paragraphs.

T3: We use the authors' implementation of the tree transducer system described in Cohn and Lapata (2008). T3 similarly requires sentence-level input/output pairs, but can also learn from abstractive compressions. We thus used a larger set of approximately 28,000 examples (single sentences with abstractive compressions taken directly from the data or as a result of heuristic sentence-level alignment of two-sentence paragraphs). We obtained parse trees using the Stanford parser (Klein and Manning, 2003), and used Jacana (Yao et al., 2013) for word alignment. The performance obtained by T3 in our experiments is substantially weaker (relative to ILP) than that reported in prior work (Cohn and Lapata, 2008). We therefore interpret this system output solely as data for evaluating automatic metrics.

NAMAS: We run the publicly available implementation of NAMAS[12] with the settings described by Rush et al. (2015). We modified the beam search algorithm to produce output with a compression ratio similar to that of the human references, since this ratio is a large factor in compression quality (Napoles et al., 2011), and systems generally perform better if allowed to produce longer output, up to the maximum length limit. We enforced output length between 50% and 75% of input length, which resulted in improved performance.

Seq2seq: We implemented the sequence-to-sequence model[13] described in Filippova et al. (2015). A deletion-based model, it uses the deletion-based subset of our training dataset and the deletion-based subset from the external data in Table 7. The encoder and decoder have three stacked LSTM layers, the hidden dimension size is 512, and the vocabulary size is 30,000. The compression rate was controlled in the same range as for the NAMAS model.

All models produce output on all inputs in the test set. For all models, we generated outputs for multi-sentence inputs by concatenating outputs for each individual sentence.[14]

4.3 Results

Overall, we consider 80 metric variants, consisting of combinations of six types of linguistic units, combined with three scoring methods (Precision, Recall, and F-measure) and four settings of single reference SINGLE-REF or three ways of scoring against multiple references MULT-ALL,MULT-MAX,MULT-PROB. Additionally, we include the standard single and multi-reference versions of BLEU-2,BLEU-3,BLEU-4, and ROUGE-L.

We compare automatic metrics to human judgements at the level of individual outputs or groups of outputs (the whole corpus). For a single output o, the human quality judgment is defined as the average assigned by up to five human raters. We denote the meaning, grammar, and combined quality values by $M(o)$, $G(o)$, and $C(o) = .5M(o) + .5G(o)$, respectively. We define the quality for a group of outputs as the arithmetic mean of judgments over the outputs in the group. We use the arithmetic mean of automating metrics at the individual output level to define automatic corpus quality metrics as well.[15] To compare different metrics and establish statistical significance of the difference between two metrics, we use Williams test of the significance of the difference

[11] https://github.com/cnap/sentence-compression

[12] https://github.com/facebook/NAMAS

[13] https://github.com/ketranm/tardis

[14] In small scale preliminary manual evaluation, we found that, although some models are theoretically able to make use of context beyond the sentence boundary, they performed better if they compressed each sentence in a sequence independently.

[15] This method has been standard for ROUGE, but has not for BLEU. We find that averaging sentence-level metrics is also advantageous for BLEU .

System	Meaning	Grammar	Combined
T3	1.14	1.40	1.26
NAMAS	1.56	1.30	1.43
Seq2Seq	1.64	1.51	1.57
ILP	2.28	2.22	2.25

Table 8: Average human ratings of system outputs for meaning and grammar separately and in combination.

between dependent Pearson correlations with human judgments (Williams, 1959) as recommended for summarization evaluation (Graham, 2015) and other NLP tasks (e.g. (Yannakoudakis et al., 2011)).

4.3.1 Corpus-level metrics

Table 8 shows the average human ratings of the four systems, separately in meaning and grammar, as well as the combined measure (an arithmetic mean of meaning and grammar judgments). Even though the performance of some systems is similar, the differences between all pairs of systems in meaning and grammar are significant $p < 0.0001$ according to a paired t-test. It is interesting to note that ILP outperforms the more recently developed neural network systems Seq2Seq and NAMAS. This might seem to contradict recent results showing that the new models are superior to traditional baselines, such as ILP. We note however that performance on the test corpus in our study might not substantially improve through the use of large automatically mined data-sets of headlines and corresponding news article sentences, due to differences in genre and domain. Using such data-sets for effective training of neural network models for non-newswire domains remains an open problem.

For each of the 80 metrics, we compared the ranking of the four systems with the ranking according to average human quality. Fifty three of the metrics achieved perfect Spearman ρ and Kendall τ_B correlation with human judgments of combined meaning and grammar quality. Due to the small sample size (four systems), we are unable to find statistically significant differences among metrics at the corpus level. We only note that precision-based metrics involving large linguistic units (four-grams) had negative correlations with human judgments. We can conclude, however, that evaluation at the corpus level is robust for a wide variety of standard metrics using linguistic units of size three or smaller.

4.3.2 Single input-level pairwise system comparisons

We can garner greater insight into the difference of metric performance when we compare metrics at the single input level. To gauge the ability of metrics to comparatively evaluate the quality of two systems, we compute single input-level correlations of automatic metrics with human judgments following the protocol of Galley et al. (2015). Each system A produces a sequence of outputs ${o_1}^A, \ldots, {o_n}^A$, corresponding to inputs $x_1, \ldots, x_n$. For each system output, we use $Q(a)$ to denote a generic human quality metric, varying over meaning, grammar, and their combination. For each pair of systems A and B, and each metric m, we compute the difference in quality for corresponding system outputs for each input x_i: $m({o_i}^A) - m({o_i}^B)$ and the difference in quality according to human judgments: $Q({o_i}^A) - Q({o_i}^B)$, and compute the correlation between these two sequences. We can thus compute the single input-level correlation between m and Q for each pair of systems A and B, resulting in a total of six correlation values (for the six pairs of systems) for each metric. For each pair of metrics m_1 and m_2, and for each pair of systems A and B, we compute the statistical significance of the difference between the Pearson correlations of these metrics with human judgements. We say that m_1 is significantly better than m_2 on the A vs. B comparison if its Pearson correlation with human quality Q is significantly better (according to the Williams test of the difference in dependent correlations) than that of m_2 with a p-value less than .05. We say that m_1 dominates m_2 overall if it is significantly better than m_2 on at least 80% of the pair-wise system comparisons.

Table 9 shows the main correlation results at the level of individual inputs. We report correlations with meaning, grammar, and combined quality separately. For each human quality metric, we see the top automatic metrics in the first group of rows. The top metrics are ones that, for at least 80% of the system comparisons, are not significantly dominated by any other metric. In addition, we show the impact of each of the three criteria: linguistic units, scoring measure, and multiple references, in corresponding groups of rows. For each linguistic unit type, we show the best-performing metric that uses units of

Top metrics					
SKIP-2+Recall+MULT-PROB	.59	PARSE-2+Recall+MULT-MAX	.35	PARSE-2+Recall+MULT-PROB	.52
PARSE-2+Recall+MULT-PROB	.57	LR-3+F-1+MULT-ALL	.35	PARSE-2+Recall+MULT-MAX	.52
SKIP-2+Recall+MULT-MAX	.58	PARSE-2+F-1+MULT-ALL	.35	SKIP-2+Recall+MULT-MAX	.51
		PARSE-2+Recall+MULT-PROB	.35	LR-2+Recall+MULT-PROB	.51
		LR-2+F-1+MULT-ALL	.34	LR-2+F-1+MULT-ALL	.50
		LR-3+Recall+MULT-ALL	.34		
Best per linguistic unit					
LR-1+Recall+MULT-PROB	.54	**LR-1**+Recall+MULT-MAX	.25^{-}	**LR-1**+Recall+MULT-PROB	.44^{-}
LR-2+Recall+MULT-PROB	.56*	**LR-2**+F-1+MULT-ALL	.34*	**LR-2**+Recall+MULT-PROB	.51*
LR-3+Recall+MULT-ALL	.55*	**LR-3**+F-1+MULT-ALL	.35*	**LR-3**+F-1+MULT-ALL	.50*
LR-4+Recall+MULT-ALL	.52^{-}	**LR-4**+F-1+MULT-ALL	.34*	**LR-4**+Recall+MULT-ALL	.47
SKIP-2+Recall+MULT-PROB	.59*	**SKIP-2**+F-1+MULT-PROB	.33	**SKIP-2**+Recall+MULT-MAX	.51*
PARSE-2+Recall+MULT-PROB	.57*	**PARSE-2**+Recall+MULT-MAX	.35*	**PARSE-2**+Recall+MULT-PROB	.52*
Best per scoring type					
SKIP-2+**Recall**+MULT-PROB	.59*	PARSE-2+**Recall**+MULT-MAX	.35	PARSE-2+**Recall**+MULT-PROB	.52
SKIP-2+**Precision**+MULT-ALL	.36^{-}	LR-2+**Precision**+MULT-ALL	.31	SKIP-2+**Precision**+MULT-ALL	.37^{-}
SKIP-2+**F-1**+MULT-ALL	.58*	LR-3+**F-1**+MULT-ALL	.35	LR-2+**F-1**+MULT-ALL	.50
Best per reference aggregation					
SKIP-2+Recall+**SINGLE-REF**	.49^{-}	PARSE-2+F-1+**SINGLE-REF**	.29	SKIP-2+Recall+**SINGLE-REF**	.44^{-}
SKIP-2+Recall+**MULT-MAX**	.58*	PARSE-2+Recall+**MULT-MAX**	.35	PARSE-2+Recall+**MULT-MAX**	.52
SKIP-2+Recall+**MULT-PROB**	.59*	PARSE-2+Recall+**MULT-PROB**	.35	PARSE-2+Recall+**MULT-PROB**	.52
SKIP-2+Recall+**MULT-ALL**	.58*	LR-3+F-1+**MULT-ALL**	.35	LR-2+F-1+**MULT-ALL**	.50
Other standard setting combinations					
BLEU-3+PrecBrev+MULT-ALL	.50^{-}	BLEU-4+PrecBrev+MULT-ALL	.30	BLEU-3+PrecBrev+MULT-ALL	.45^{-}
ROUGE-L+Recall+MULT-MAX	.49^{-}	ROUGE-L+Recall+MULT-MAX	.27	ROUGE-L+Recall+MULT-MAX	.43

Table 9: Left to right: Pearson correlation of automatic metrics with human ratings for meaning, grammar, and combined quality.

this type. Similarly, for the other criteria, we show the best performing metric for each value of the criterion. Metrics with a * suffix in each group significantly dominate metrics with a $^{-}$ suffix. Metrics with a $^{-}$ suffix in a group are dominated by at least one other metric, possibly outside of the group. The lowest group of rows in each main column presents the performance of other metrics that cannot be classified directly based on the three criteria.

A high-level observation that can be made is that the correlations with meaning are much higher than the correlations with grammar. The best correlations in meaning can be classified as "strong", whereas the best correlations in grammar are in the "medium" range. Unigrams are heavily dominated by higher order n-grams in all settings. Four-grams are also weaker that other units in measuring meaning preservation. Dependency triple (parse-based) metrics are strong, in particular in measuring grammaticality, but do not significantly dominate skip bi-grams or contiguous bi-grams. The scoring measure used has a strong impact. We see that precision-based metrics are substantially dominated by metrics that incorporate recall, except for grammar evaluation. Importantly, we see that multiple references contribute substantially to metric quality, as all methods that use multiple references outperform single-reference metrics. In both meaning and combined evaluation, this difference was statistically significant. Finally, we observe that standard BLEU metrics and ROUGE-L were not competitive.

5 Conclusion

We have introduced a large manually collected multi-reference abstractive dataset and quantified the impact of editing operations and context on human compression quality, showing that substitution and rephrasing operations are more meaning preserving than other operations, and that compression in context improves quality. Further, in the first systematic study of automatic evaluation metrics for text compression, we have demonstrated the importance of utilizing multiple references and suitable linguistic units, and incorporating recall.

Acknowledgments

We are grateful to Jaime Teevan, Shamsi Iqbal, Dan Liebling, Bill Dolan, Michel Galley, and Wei Xu, together with the three anonymous reviewers for their helpful advice and suggestions.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG*.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of ACL-HLT*.

Ted Briscoe and John A Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC*.

Hori Chiori and Sadaoki Furui. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems*, 87(1):15–25.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of ACL-COLING*.

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

James Clarke. 2008. *Global Inference for Sentence Compression: An Integer Linear Programming Approach*. Ph.D. thesis, Univeristy of Edinburgh.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of EMNLP*.

Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of EMNLP*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of ACL-IJCNLP (Volume 2: Short Papers)*.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of EMNLP*.

Mingkun Gao, Wei Xu, and Chris Callison-Burch. 2015. Cost optimization in crowdsourcing translation: Low cost translations made even cheaper. In *Proceedings of NAACL-HLT*.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of EMNLP*.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, Charles J. Fillmore, and Rebecca J. Passonneau. 2008. MASC: the manually annotated sub-corpus of american english. In *Proceedings of LREC*.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of ACL (Volume 2: Short Papers)*.

Hongyan Jing and Kathleen R McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of SIGIR*.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of ANLP*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. 2011. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*.

Ellie Pavlick and Joel Tetrault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, pages 61–74.

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 NAACL-HLT*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP*.

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 COLING-ACL*.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Evan James Williams. 1959. *Regression analysis*. Wiley, New York.

Wei Xu, Courtney Napoles, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of ACL-HLT*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of ACL (Volume 2: Short Papers)*.