Paraphrastic Sentence Compression with a Character-based Metric: Tightening without Deletion

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

Abstract

We present a substitution-only approach to sentence compression which "tightens" a sentence by reducing its character length. Replacing phrases with shorter paraphrases yields paraphrastic compressions as short as 60% of the original length. In support of this task, we introduce a novel technique for re-ranking paraphrases extracted from bilingual corpora. At high compression rates[1] paraphrastic compressions outperform a state-of-the-art deletion model in an oracle experiment. For further compression, deleting from oracle paraphrastic compressions preserves more meaning than deletion alone. In either setting, paraphrastic compression shows promise for surpassing deletion-only methods.

1 Introduction

Sentence compression is the process of shortening a sentence while preserving the most important information. Because it was developed in support of extractive summarization (Knight and Marcu, 2000), much of the previous work considers deletion-based models, which extract a subset of words from a long sentence to create a shorter sentence such that meaning and grammar are maximally preserved. This framework imposes strict constraints on the task and does not accurately model human-written compressions, which tend to be abstractive rather than extractive (Marsi et al., 2010).

We distinguish two non-identical notions of sentence compression: making a sentence substantially shorter to conform to a stated maximum length versus "tightening" a sentence by removing unnecessary verbiage. We propose a method to tighten sentences with no deletion operations, just substitution. Using paraphrases extracted from bilingual text and re-ranked on monolingual data, our system selects the set of paraphrases that minimizes the character length of a sentence.

While not currently the standard, character-based lengths have been considered before in compression, and we believe that it is relevant for current and future applications. Character lengths were used for document summarization (DUC 2004, Over and Yen (2004)), summarizing for mobile devices (Corston-Oliver, 2001), and subtitling (Glickman et al., 2006). Although in the past strict word limits were often imposed for various documents, information transmitted electronically is limited by the number of bytes, which directly relates to number of characters. Mobile devices, SMS messages, and microblogging sites such as Twitter are increasingly important for quickly spreading information. In this context, it is important to consider character-based constraints.

Twitter is increasingly popular for sharing information quickly. Character-based compression allows more information to be conveyed in 140 characters (the length constraint of Twitter posts or *tweets*). For example, many article lead sentences exceed this limit. A paraphrase substitution oracle compresses the sentence in the table below to 76% of its original length (162 to 123 characters; the first is the original).[2] With a 17-character shortened link to the article, it is 140 characters including spaces.

[1] Compression rate is defined as the compression length over original length, so lower values indicate shorter sentences.

[2] Taken from the main page of http://wsj.com, April 9, 2011.

Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.

Congress made a final agreement Fri. to avoid government shutdown, after days of haggling and tense hours of brinkmanship. on.wsj.com/h8N7n1

In contrast, using deletion to compress to the same length may not be as expressive:

Congressional leaders reached agreement Friday to avert a shutdown of federal government, after haggling and tense hours. on.wsj.com/h8N7n1

This work presents a model that makes paraphrase choices to minimize the *character* length of a sentence. Even with recent innovations in paraphrasing, unsuitable paraphrase choices are still present. An oracle paraphrase substitution experiment shows that human judges rate paraphrastic compressions higher than deletion-based compressions. To achieve further compression, we shortened the oracle compressions using a deletion model to yield compressions 80% of the original sentence length and compared these to compressions generated using just deletions. Humans found the oracle-then-deletion compressions to preserve more meaning than deletion-only compressions at uniform compression rates.

2 Related work

Most of the previous work focuses on deletion using syntactic information, e.g. (Galley and McKeown, 2007; Knight and Marcu, 2002; Nomoto, 2009; Galanis and Androutsopoulos, 2010; Filippova and Strube, 2008; McDonald, 2006; Yamangil and Shieber, 2010; Cohn and Lapata, 2008; Cohn and Lapata, 2009; Turner and Charniak, 2005). Woodsend et al. (2010) incorporate paraphrase rules in a deletion model. Previous work in subtitling has made one-word substitutions to decrease character length at high compression rates (Glickman et al., 2006). More recent approaches in steganography have used paraphrase substitution to encode information in text but focus on grammaticality, not meaning preservation (Chang and Clark, 2010).

Sentence compression has been considered before in contexts outside of summarization, such as headline, title, and subtitle generation (Dorr et al., 2003; Vandeghinste and Pan, 2004; Marsi et al., 2009). Zhao et al. (2009) applied an adaptable paraphrasing pipeline to sentence compression, optimizing for F-measure over a manually annotated set of gold standard paraphrases. Corston-Oliver (2001) deleted characters from words to shorten the character length of sentences. To our knowledge character-based compression has not been examined before with the surging popularity and utility of Twitter.

3 Sentence Tightening

The distinction between tightening and compression can be illustrated by considering how much space needs to be preserved. In the case of microblogging, often a sentence has a few too many characters and needs to be "tightened". On the other hand, if a sentence is much longer than a desired length, more drastic compression is necessary. The first subtask is relevant in any context with strict word or character limits. Some sentences may not be compressible beyond a certain limit, for example we found that near 10% of the compressions generated by Clarke and Lapata (2008) were identical to the original sentence. In situations where the sentence *must* meet a minimum length, tightening can be used to meet these requirements.

Multi-reference translation provide an instance of the natural length variation of human-generated sentences. The translations represent different ways to express the same sentence, so there should be no meaning lost between the reference translations. The character-based length of different translations of a given sentence varies on average by 80% when compared to the shortest sentence in a set.[3] This provides evidence in favor of tightening a sentence to some extent without losing any meaning.

Through the lens of sentence tightening, we consider whether paraphrase substitutions alone can yield compressions competitive with a deletion at the same length. A character-based compression rate is crucial in this framework, as two compressions having the same *character* compression rate may have different *word-based* compression rates. The advantage of a character-based substitution model is in choosing shorter words when possi-

[3] This value will vary by collection and with the number of references: for example, the NIST05 Arabic reference set has a mean ratio of 0.92 with 4 references per set.

ble, creating space to preserve more content words. This framework could be limited to consider only paraphrases with fewer words than the original, but there is no guarantee that the new paraphrase will have fewer characters. Indeed, paraphrases with the same number of words (or more) as the original phrase frequently have fewer characters.

3.1 Paraphrase Acquisition

To generate paraphrases for use in our experiments, we took the approach described by Bannard and Callison-Burch (2005), which extracts paraphrases from bilingual parallel corpora. Figure 1 illustrates the process. A phrase to be paraphrased, like *thrown into jail*, is found in a German-English parallel corpus. The corresponding foreign phrase (*festgenommen*) is identified using word alignment and phrase extraction techniques from phrase-based statistical machine translation (Koehn et al., 2003). Other occurrences of the foreign phrase in the parallel corpus may align to another English phrase like *jailed*. Following Bannard and Callison-Burch (2005), we treated any English phrases that share a common foreign phrase as potential paraphrases of each other.

As the original phrase occurs several times and aligns with many different foreign phrases, each of these may align to a variety of other English paraphrases. Thus, *thrown into jail* not only paraphrases as *jailed*, but also as *arrested*, *detained*, *imprisoned*, *incarcerated*, *locked up*, *taken into custody*, and *thrown into prison* and others like *be thrown in prison*, *been thrown into jail*, *being arrested*, *in jail*, *in prison*, *put in prison for were thrown into jail*, and *who are held in detention*. Moreover, because the method relies on noisy and potentially inaccurate word alignments, it is prone to generating many bad paraphrases, such as *maltreated*, *thrown*, *cases*, *custody*, *arrest*, *owners*, and *protection*.

To rank candidates, Bannard and Callison-Burch defined the paraphrase probability $p(e_2|e_1)$ based on the translation model probabilities $p(e|f)$ and $p(f|e)$ from statistical machine translation. Following Callison-Burch (2008), we refine selection by requiring both the original phrase and paraphrase to be of the same syntactic type, which leads to more grammatical paraphrases.

Although many excellent paraphrases are extracted from parallel corpora, many others are unsuitable and the translation score does not always accurately distinguish the two. Therefore, we reranked our candidates based on monolingual distributional similarity, employing the method described by Van Durme and Lall (2010) to derive approximate cosine similarity scores over feature counts using single token, independent left and right contexts. Features were computed from the web-scale n-gram collection of Lin et al. (2010). As 5-grams are the highest order of n-gram in this collection, this constrained the set of paraphrases to be at most length four (which allows at least one word of context).

Paraphrase	**Monlingual**	**Bilingual**
study in detail	1.00	0.70
scrutinise	0.94	0.08
consider	0.90	0.20
keep	0.83	0.03
learn	0.57	0.10
study	0.42	0.07
studied	0.28	0.01
studying it in detail	0.16	0.05
undertook	0.06	0.06

Table 1: Candidate paraphrases for *study in detail* with corresponding approximate cosine similarity (Monolingual) and translation model (Bilingual) scores.

To our knowledge this is the first time such techniques have been used in combination, in order to derive higher quality paraphrase candidate pairs. See Table 1 for an example.

The monolingual-filtering we describe is by no means limited to paraphrases extracted from bilingual corpora. It could be applied to other data-driven paraphrasing techniques (see Madnani and Dorr (2010) for a survey). Although it is particularly well suited to the bilingual extracted corpora, since the information that it adds is orthogonal to that model, it would presumably add less to paraphrasing techniques that already take advantage of monolingual distributional similarity (Pereira et al., 1993; Lin and Pantel, 2001; Barzilay and Lee, 2003).

In order to evaluate the paraphrase candidates and scoring techniques, we randomly selected 1,000 paraphrase sets where the source phrase was present in the corpus described in Clarke and Lapata (2008). For each phrase and set of candidate paraphrases, we extracted all of the contexts from the corpus in which the source phrase appeared. Human judges were

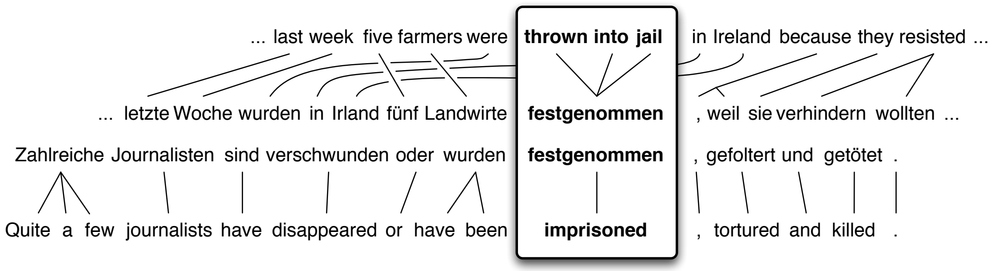
... last week five farmers were
thrown into jail
in Ireland because they resisted ...
... letzte Woche wurden in Irland fünf Landwirte
festgenommen
, weil sie verhindern wollten ...
Zahlreiche Journalisten sind verschwunden oder wurden
festgenommen
, gefoltert und getötet .
Quite a few journalists have disappeared or have been
imprisoned
, tortured and killed .

Figure 1: Using a bilingual parallel corpus to extract paraphrases.

presented with the original sentences and then the same sentences with each paraphrase candidate substituted in. Each paraphrase substitution was graded based on the extent to which it preserved the meaning and how much it affected the grammaticality of the sentence. While both the bilingual translation score and monolingual cosine similarity positively correlated with human judgments, the monolingual score proved a stronger predictor of quality in both dimensions. Using Kendall's tau correlation coefficient, the agreement between the ranking imposed by the monolingual score and human ratings surpassed that of the original ranking as derived during the bilingual extraction, for both meaning and grammar.[4] In our substitution framework, we ignore the translation probabilities and use only the approximate cosine similarity in the paraphrase decision task.

4 Framework for Sentence Tightening

Our sentence tightening approach uses a dynamic programming strategy to find the combination of non-overlapping paraphrases that minimizes a sentence's character length. The threshold of the monolingual score for paraphrases can be varied to widen or narrow the search space, which may be further increased by considering any lexical paraphrases not subject to syntactic constraints. Sentences with a compression rate as low as 0.6 can be generated without thresholding the paraphrase scores. Because the system can generate multiple paraphrased sentences of equal length, we apply two layers of filtering to generate a single output. First we calculate a word-overlap score between original and candidate sentences to favor compressions similar to the original sentence; then, from among the sentences with the highest word overlap, we select the compression with the best language-model score.

Higher paraphrase thresholds guarantee more appropriate paraphrases but yield longer compressions. Using a threshold of 0.95, the average compression rate is 0.968, which is considerably longer than the compressions using no threshold (0.60). In these experiments we did not syntactically constrain paraphrases.

In case where judges favor compressions with high word overlap with the original sentence, we compressed the longest sentence from each set of reference translations (Huang et al., 2002) and randomly chose a sentence from the set of reference translations to use as the standard for comparison. Paraphrastic compressions were generated at paraphrase-score thresholds ranging from 0.60 to 0.95. We implemented a state-of-the-art deletion model (Clarke and Lapata, 2008) to generate deletion-only compressions. We fixed the compression length to ± 5 characters of the length of each paraphrastic compression, in order to isolate the compression quality from the effect of compression rate. The experiments were done using Amazon's Mechanical Turk with three-way redundancy and two 5-point scales for meaning and grammar (5 being the highest score).

deleted citation to mturk wkshp

5 Evaluation

The initial results of our substitution system show room for improvement in future work (Table 2). We believe this is due to erroneous paraphrase substitutions, since phrases with the same syntactic cate-

[4] For meaning and grammar respectively, $\tau = 0.28$ and 0.31 for monolingual scores and 0.19 and 0.15 for bilingual scores.

System	Grammar	Meaning	CompR	Cos.
Substitution	3.8	3.7	0.97	0.95
Deletion	4.1	4.0	0.97	-
Substitution	3.4	3.2	0.89	0.85
Deletion	4.0	3.8	0.89	-
Substitution	3.1	3.0	0.85	0.75
Deletion	3.9	3.7	0.85	-
Substitution	2.9	2.9	0.82	0.65
Deletion	3.8	3.5	0.82	-

Table 2: Mean ratings of compressions using just deletion or substitution at different paraphrase thresholds (Cos). Deletion performed better in all settings.

gory and distributional similarity are not necessarily semantically identical. Illustrative examples include *WTO* for *United Nations* and *east* or *west* for *south*. The quality of the multi-reference translations is not uniformly high, so we used a dataset of English newspaper articles for the following experiment.

To control against these errors and test the viability of a substitution-only approach, we generated all possible paraphrase substitutions above a threshold of 0.80 within a set of 20 randomly chosen sentences from the written corpus of Clarke and Lapata (2008). We solicited humans to make a ternary decision of whether a paraphrase was acceptable in the context (*good*, *bad*, or *not sure*), and generated oracle compressions using only paraphrase substitutions on which all three annotators agreed that the paraphrase was *good*.

Employing the deletion model, we generated compressions constrained to ± 5 characters of the length of the oracle compression. The oracle generated compressions with an average compression rate of 0.90. Next, we examined whether applying the deletion model to paraphrastic compressions would improve compression quality. In manual evaluation along the dimensions of grammar and meaning, both the oracle compressions and oracle-plus-deletion compressions outperformed the deletion-only compressions at uniform lengths (Table 3)[5]. These results suggest that improvements in paraphrase acquisition will make our system competitive with deletion-only models.

Model	Grammar	Meaning	CompR
Oracle	4.1	4.3	0.90
Deletion	4.0	4.1	0.90
Gold	4.3	3.8	0.75
Oracle+deletion	3.4	3.7	0.80
Deletion	3.2	3.4	0.80

Table 3: Mean ratings of compressions generated by a substitution oracle, deletion only, deletion on the oracle compression, and the gold standard. Being able to choose the best paraphrases would enable our substitution model to outperform the deletion model.

6 Conclusion

This work shows promise for the use of only substitution in the task of sentence tightening. There are myriad possible extensions and improvements to this method, most notably richer features beyond paraphrase length. We do not currently use syntactic information in our paraphrastic compression model because it places limits on the number of paraphrases available for a sentence and thereby limits the possible compression rate. However, we believe that our monolingual refining of paraphrase sets improves paraphrase selection and is a reasonable alternative to using syntactic constraints. The current method for paraphrase extraction does not include certain types of rewriting, such as passivization, and should be extended to incorporate even more shortening paraphrases. Future work can directly apply these methods to Twitter and extract additional paraphrases and abbreviations from Twitter and/or SMS data. Our substitution approach can be improved by applying different techniques to choosing the best candidate compression, or by framing it as an optimization problem over more than just minimal length. Overall, we find these results to be encouraging for the possibility of sentence compression without deletion.

Acknowledgments

We are grateful to John Carroll for helping us obtain the RASP parser. This research was partially funded 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.

[5] Paraphrastic compressions were rated significantly higher for meaning, $p < 0.05$

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*.

Ching-Yun Chang and Stephen Clark. 2010. Linguistic steganography using automatically generated paraphrases. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 591–599. Association for Computational Linguistics.

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.

Simon Corston-Oliver. 2001. Text compaction for display on very small screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*.

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*.

Katja Filippova and Michael Strube. 2008. Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference*. Association for Computational Linguistics.

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*.

Oren Glickman, Ido Dagan, Mikaela Keller, Samy Bengio, and Walter Daelemans. 2006. Investigating lexical substitution scoring for subtitle generation. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 45–52. Association for Computational Linguistics.

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

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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules from text. *Natural Language Engineering*, 7(3):343–360.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New Tools for Web-Scale N-grams. In *Proceedings of LREC*.

Nitin Madnani and Bonnie Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–388.

Erwin Marsi, Emiel Krahmer, Iris Hendrickx, and Walter Daelemans. 2009. Is sentence compression an NLG task? In *Proceedings of the 12th European Workshop on Natural Language Generation*.

Erwin Marsi, Emiel Krahmer, Iris Hendrickx, and Walter Daelemans. 2010. On the limits of sentence compression by deletion. *Empirical Methods in Natural Language Generation*, pages 45–66.

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

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

Paul Over and James Yen. 2004. An introduction to DUC 2004: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2004 Document Understanding Workshop, Boston*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *ACL-93*.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of ACL*.

Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *Proceedings of ACL, Short Papers*.

Vincent Vandeghinste and Yi Pan. 2004. Sentence compression for automated subtitling: A hybrid approach. In *Proceedings of the ACL workshop on Text Summarization*.

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

Elif Yamangil and Stuart M. Shieber. 2010. Bayesian synchronous tree-substitution grammar induction and its application to sentence compression. In *Proceedings of ACL*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation.