Springer Link

Search

International Conference on Intelligent Text Processing and Computational Linguistics
CICLing 2014: Computational Linguistics and Intelligent Text Processing pp 553-565 | Cite as

Evaluation of Sentence Compression Techniques against Human Performance

Authors Authors and affiliations

Prasad Perera, Leila Kosseim

Part of the Lecture Notes in Computer Science book series (LNCS, volume 8404)

Abstract

Evaluation of Sentence Compression Techniques against Human Performance

Prasad Perera and Leila Kosseim

Dept. of Computer Science & Software Engineering
Concordia University
Montreal, Canada
{p_perer,kosseim}@encs.concordia.ca

Abstract. This paper presents a comparison of various sentence compression techniques with human compressed sentences in the context of text summarization. Sentence compression is useful in text summarization as it allows to remove redundant and irrelevant information hence preserve space for more relevant information. In this paper, we evaluate recent state-of-the-art sentence compression techniques that are based on syntax alone, a mixture of relevancy and syntax, part of speech feature based machine learning, keywords alone and a naïve random word removal baseline. Results show that syntactic based techniques complemented by relevancy measures outperform all other techniques to preserve content in the task of text summarization. However, further analysis of human compressed sentences also shows that human compression techniques rely on world knowledge which is not captured by any automatic technique.

1 Introduction

The goal of sentence compression is to generate a more concise form of a sentence without losing its grammaticality or its relevant content. Sentence compression has been used in several downstream natural language processing applications such as text simplification [1], headline generation [2] and text summarization [3]. In extractive summarization, relevant sentences are extracted from the document collection and re-ordered to create the final summary. However, since these sentences are not processed or modified, they might contain irrelevant content or phrases that do not contribute much to the summary content. As an example, consider the following topic, query and sentence (1)[1]:

Topic: *Microsoft's antitrust problems*

Query: *Summarize Microsoft's antitrust problems, including its alleged illegal behavior and antitrust proceedings against the company*

(1) *Under the schedule set by Jackson in April, the Justice Department and 17 states filed a brief with the court on April 28 asking the judge to break Microsoft into two companies as the remedy for the illegal behavior found in the long antitrust trial.*

[1] All examples are taken from the DUC 2007 corpora.

A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 553–565, 2014.

The candidate sentence (1) does contain relevant pieces of information, but it may be too long to fit into a summary as it is. In addition, the sentence also contains several pieces of information that are less relevant to the topic and the query given and if it were to be inserted as in, it might lower the relevancy of the overall summary. A few possible shorter or compressed forms of the above sentence include:

> (1c1) *~~Under the schedule set by Jackson in April,~~ the Justice Department and 17 states filed a brief with the court on April 28 asking the judge to break Microsoft into two companies as the remedy for the illegal behavior found in the long antitrust trial.*
>
> (1c2) *Under the schedule set by Jackson in April, the Justice Department and 17 states filed a brief with the court ~~on April 28~~ asking the judge to break Microsoft into two companies as the remedy ~~for the illegal behavior found in the long antitrust trial~~.*
>
> (1c3) *~~Under the schedule set by Jackson in April,~~ the Justice Department and 17 states filed a brief with the court ~~on April 28~~ asking the judge to break Microsoft into two companies as the remedy ~~for the illegal behavior found in the long antitrust trial~~.*

To the best of our knowledge, very little previous work has focused on measuring the contribution of specific sentence compression techniques as a means to improve summary content.

2 Sentence Compression Techniques

In this section, we present an overview of previous sentence compression techniques that have been proposed. These techniques are based on syntactic structure based pruning, keyword based pruning, part of speech and syntax related features based machine learning, or a combination of syntax and relevancy based pruning. The following sections present these in detail.

2.1 Syntactic Pruning

Predefined fixed syntactic pruning is the basis of many sentence compression techniques. For example, [4] used complete dependency parses and pruned specific grammatical structures including prepositional complements of verbs, subordinate clauses, noun appositions and interpolated clauses. They achieved a compression rate of 26% while retaining grammaticality and readability of text. In [5], the authors also applied linguistically motivated syntactic filtering. Using the TIPSTER [6] corpus, they identified syntactic patterns which were absent from human-written summaries compared to the original corpus and defined a trimming algorithm consisting of removing sub-trees of specific grammatical phrase structures. They have evaluated their pruning technique on the DUC 2003 summarization task and showed an improvement in ROUGE scores compared to uncompressed length-limited summaries. Finally, [7] describes the sentence compression module of their text summarization system, based on syntactic level sentence pruning. They have implemented a module of compression which filters adverbial modifiers and relative clauses. Their evaluations, performed using the DUC 2007 summarization track, showed an improvement in ROUGE scores after applying their compression technique to their summarization system.

2.2 Machine Learning Techniques

Machine learning techniques differ from fixed syntactic pruning in that the words, phrases or syntactic structures to prune are learned automatically from annotated corpora. [8, 9], for example, remove inessential phrases in extractive summaries based on an analysis of human written abstracts. A syntactic parser is used to identify different types of phrases which are present in the original sentences but not in human written simplified sentences. These phrases along with the main noun or verb they are attached to are used to train a Naïve Bayes Classifier to decide how likely a phrase is to be removed from a sentence. For evaluation, they have used a metric called success rate, which computes the ratio between the number of correct prunings the classifier made that agree with the human annotations over the total number of prunings (the classifier made and the humans made). The authors have achieved a 78.1% overall success rate but have noted a low success rate in removing adjectives, adverbs and verb phrases. However, the effect on overall summary content was not indicated.

On the other hand, [10] introduced semantic features to improve a decision tree based classification. Here, the authors used Charniak's parser [11] to generate syntactic trees and incorporated semantic information using WordNet [12]. The evaluation showed a slight improvement in importance of information preserved in shortened sentences. However again, the effect on summarization was not noted.

[13] points out that text compression could be seen as a problem of finding a global optimum by considering the compression of the whole text/document. The authors used syntactic trees of each pair of long and short sentences to define rules to deduce shorter syntactic trees out of original syntactic trees. They used the Ziff-Davis corpus for their evaluation as well as human judgment. Similarly, [14] describes the use of integer linear programming model to infer globally optimal compressions while adhering to linguistically motivated constraints and show improvement in automatic and human judgment evaluations. [15] have also described a syntactic pruning approach based on transformed dependency trees and a linear integer model.

[16] described another machine learning approach to sentence compression. They first trained a probabilistic model based on Maximum Entropy (ME) to evaluate how likely an edge of a syntactic tree can be removed based on a set of features including the part of speech tags of the surrounding words of the edge, the head of the edge and the modifier of the edge. For the evaluation, summaries were judged by human annotators and results showed that their compression

techniques outperform the baseline algorithm yet underperformed compared to the human annotated compressions.

2.3 Hybrid Methods

The sentence compression technique presented in [17] is based on two main parameters: syntactic pruning heuristics and a relevancy score. Based on these two parameters, three different sentence compression techniques were defined:

1. Syntax-driven sentence pruning: The goal of this technique is to preserve the grammaticality of the sentences while removing predefined syntactic structures that are assumed to always contain secondary information. Hence it is expected that removing these grammatical structures would not remove relevant content significantly. Based on the work of [5] and [4] (adapted to English), six types of syntactic structures are pruned: Relative clauses, Adjective phrases, Adverbial phrases, Conjoined verb phrases, Appositive phrases, Prepositional phrases. This syntax pruning method is equivalent to those presented in Section 2.1.

2. Syntax with relevancy based pruning: This method tries to ensure that the syntactic structures pruned do not contain relevant information. [17] use a relevance score based on tf-idf and cosine similarity with the topic/query. A relevance score for each sub-tree is calculated and predefined syntactic sub-trees are pruned only if relevance score is smaller than a given threshold.

3. Relevancy-driven syntactic pruning: This technique focuses primarily on the relevance score of syntactic structure. Here, no predefined syntactic structures are used. A relevancy score is computed for each syntactic substructure and the lowest embedding syntactic structures are removed if their relevance score is below a given threshold. For the sake of preserving the grammaticality of resulting sentences, syntactic sub-structures that are marked as noun or verb phrases are not removed.

2.4 Keyword/Phrase Based Techniques

Keyword based techniques take a more conservative approach and use a predefined list of keywords or phrases to identify less significant parts of the text and remove them from long sentences. [18–20], for example used a keyword list implemented in an ad hoc fashion to omit specific terms. They have evaluated their pruning techniques within their CLASSY [19] summarization system with DUC 2005 [21] and showed an improvement in ROUGE scores. In their participation to the DUC 2006 [22] automatic summarization track, their system scored among the top three based on ROUGE scores. The goal of our work was to apply each type of sentence compression technique for the task of summarization and evaluate them against human compression techniques. To do so, we have performed a series of evaluations based on two factors: the compression rates and content evaluation of compressed summaries. For content evaluation, we have evaluated ROUGE measures [23] on compressed summaries and grammar structure overlap [24, 25, 15] between human compressed summaries and summaries compressed with the automatic sentence pruning techniques. Specifically, we evaluated the following techniques:

Keyword/Phrase Based Approach. We implemented a keyword based sentence compression using the word/phrase patterns described in [18] and [20] plus additional patterns that we learned by analyzing human annotated summaries. The particular keyword/phrase and patterns we used include: meta-data information, temporal words/phrases, attributive words/phrases, keywords and key phrases (e.g. *As a result*, *In contrast*, etc.) and specific clauses (e.g. clauses that starts with *which*, *where* or *whom*, etc.).

Machine Learning Approach. As the second approach, we chose the sentence compression system described in [16]. This publicly available sentence compressor[2] uses a set of features based on part of speech (POS) tags, a Maximum Entropy based classifier to prune sentences and a Support Vector Regression Model to select the best candidates of all reduced sentences. The system uses the Edinburgh's Written and Spoken corpus[3] as the training set of original sentences paired with their human written compressed sentences. Additionally, they used a language model created using about 4.5 million sentences taken from TIPSTER corpus [6] in order to rank possible compressed sentences and select the most likely one.

Syntax and Relevancy. Here, we used the work of [17] (see Section 2.3) that used syntactic pruning combined with relevancy measures. The implementation uses the Stanford Parser [26] to generate complete syntactic trees of the sentences. We have evaluated the syntactic pruning based on: syntax alone, syntax with relevancy and relevancy alone.

Baseline Compression Techniques. In order to compare these sentence compression techniques, we implemented a baseline technique that randomly removes words and phrases from summaries to reach a particular compression rate. Using this baseline, we have created compressed summaries with compression rates of 10.5%, 21.5%, 23.4%, 29.8% and 43.1% to correspond to the compression rates of each sentence pruning techniques: keyword based (10.5%), human compression and machine learning based (21.5%), syntax with relevancy based (23.4%), relevancy-driven (29.8%) and syntax-driven (43.1%) (See Section 3.2 below).

2.5 Human Compressed Summaries

To build our human compressed summaries, we have provided a set of summaries to five human annotators and asked them to reduce their length while preserving

[2] `http://nlp.cs.aueb.gr/software.html`

[3] `http://jamesclarke.net/research/resources`

important content. We provided the evaluators with a set of summaries created using the DUC 2007 summarization task [27] along with the relevant topic and the query used to create these summaries. For this task, we have chosen the summaries created by the best performing system [28] (based on ROUGE measures) at the DUC 2007 summarization track. The human annotators were asked to compress these summaries by removing words or phrases from the sentences that they considered not relevant to the given topic/query. Each sentence was to be considered independently of the others; hence the annotators could not use the context to influence their compression strategies. Human annotators were chosen from a group of undergraduate and graduate students in different science and engineering streams.

2.6 Human Compression Rate

First we evaluated the compression rates of the human compressed summaries and compared these with the compression rates achieved by the automatic pruning techniques. Table 1 shows these word-based compression rates. According to Table 1, the highest compression rate was achieved by the syntax-driven technique. The relevancy-driven technique and the syntax with relevancy based techniques achieved the next highest compressions. Although human compression varies from 18% to 25%, the annotator average (21.5%) seems to be similar to the syntax with relevancy based technique (23.4%) and the machine learning technique (21.5%). In addition, the lowest compression rate was achieved by the keyword based technique (10.5%).

Table 1. Sentence Compression Rates of Different Techniques

Technique	**No. of Words**	**Compression**
Original Summary	6237	0.0%
Keyword Based Pruning	5579	10.5%
Annotator 1	5106	18.1%
Annotator 2	5052	19.0%
Machine Learning Technique	4914	21.5%
Annotator 3	4897	21.5%
Annotator 4	4889	21.6%
Syntactic with Relevancy	4779	23.4%
Annotator 5	4657	25.3%
Relevancy-Driven	4381	29.8%
Syntax-Driven	3552	43.1%

2.7 Content Evaluation Using ROUGE

In automatic text summarization, the most standard metric used in measuring summary content is the ROUGE measure [23]. Since we applied sentence compression to text summarization, we measured the effect of the various techniques on content using the ROUGE measure, when sentence compression was applied to text summarization. Therefore our first attempt at content evaluation was based on ROUGE F-measure scores (R-2 and SU4) for the original summaries, the five sets of human compressed summaries and all other automatically compressed summaries. As Table 2 shows, there is a decrease in ROUGE-2 score between the original summaries and the human compressed summaries. On average, the annotators have a ROUGE-2 score of 0.120 and ROUGE-SU4 of 0.172 while the original summaries have a ROUGE-2 score of 0.127 and ROUGE-SU4 of 0.179. The one-tailed t-test shows that for all the annotators (except for annotator 2), the difference between ROUGE scores compared to the original summary score is statistically significant with a confidence level of 95%. The t-test identified four clusters based on ROUGE-2 scores. The first cluster contains the techniques that scored the best ROUGE-2 scores (the original summaries, annotator 2 and the keyword based technique). Compared to the original ROUGE-2 score, the keyword based technique (ROUGE-2: 0.124) does not show a significant decrease in content according the one-tailed t-test with a confidence level of 95%. However, recall from Table 1 that this technique only removes 10.5% of words while other techniques remove up to 20-40% of words. It is therefore not surprising that its ROUGE score be so high.

In the second cluster, we have all other annotators, with a ROUGE-2 score ranging from 0.119 to 0.118. These ROUGE-2 scores are significantly lower than the original summary ROUGE-2 scores. This is somewhat surprising as we would have expected the annotators not to remove too much relevant content; yet considering that on average, they removed 21.5% of the words (see Table 1).

Table 2. Content Evaluation of Compressed Summaries

Technique	**R-2**	**R-SU4**
Original Summaries	0.127	0.179
Annotator 2	0.125	0.176
Keyword Based	0.124	0.176
Annotator 1	0.119	0.172
Annotator 3	0.119	0.171
Annotator 4	0.119	0.173
Annotator 5	0.118	0.170
Baseline: Random Compression 10.5%	0.116	0.172
Syntax with Relevancy Based	0.110	0.164
Machine Learning Based	0.107	0.162
Relevancy-Driven	0.106	0.154
Baseline: Random Compression 21.5%	0.102	0.163
Baseline: Random Compression 23.4%	0.100	0.164
Baseline: Random Compression 29.8%	0.085	0.150
Syntax-Driven	0.084	0.134
Baseline: Random Compression 43.1%	0.072	0.137

They were bound to remove some content. Also surprisingly, this cluster contains the random compression 10.5% technique with a ROUGE-2 measure of 0.116.

The third cluster includes: syntax with relevancy based, machine learning, relevancy-driven and two baseline compression techniques (random compression 21.5% & 23.4%). Again, these techniques show significantly lower ROUGE-2 scores than the original summary and the average human ROUGE scores. However, when tested for significance across each other, the ROUGE-2 scores are not significantly different.

The last cluster includes the rest of the techniques: baseline random compression 29.8%, syntax-driven and baseline random compression 43.1% that scored the lowest ROUGE scores. The ranking of the techniques is more or less the same when ROUGE-SU4 scores were used for the task. When words are randomly removed, it hurts the grammaticality and the content of the summaries. However, since ROUGE is only calculated based on bi-gram co-occurrences, it justifies how random removals (10.5%, 21.5%, 23.4% and 29.8%) showed better results than most of the automatic sentence compression techniques.

2.8 Content Evaluation Based on Grammatical Relations

Previous work on sentence compression evaluations have typically focused on two different evaluation methods: ranking of compressed sentences by human judgment or evaluation against human compressed sentences by measuring content overlap. To follow recent trend (e.g. [25, 15]), we have also evaluated the sentence compression techniques and the human annotators based on a metric that takes grammatical relations into account. This metric was first introduced in [24] for automatic summary evaluation with the goal of improving automatic evaluation techniques while taking semantic information into account. The authors argue that it is easy to enhance automatic summary evaluation when a dependency parser is available by counting co-occurrences of dependency grammar structures between the gold standard summaries and automatic summaries as opposed to counting n-gram word co-occurrences as ROUGE does. This technique was used by [25] to evaluate their sentence compression technique. Following them, [15] have also used the same mechanism to evaluate their sentence compression techniques, comparing their results to the work of [25]. Table 3 shows the F-measure calculated for all techniques and all five annotators. The dependency grammar structure F-measure seems to show some interesting results. Here, we observe four clusters of grammar structure F-measure (created based on the one-tailed t-test). The first cluster clearly shows better content evaluation results for all annotators. The second cluster includes: syntax with relevancy and keyword based techniques. Here, unlike the ROUGE measure (in Table 2), the dependency grammar metric has penalized the keyword based technique compared to the human summaries. The third cluster includes machine learning based, relevancy-driven, syntax-driven and random compression (10.5%) techniques. Finally, cluster four includes, the rest of the random removals techniques.

Using this grammar-based content metric, all baseline techniques, where words are removed randomly, have been penalized as expected. In addition,

Table 3. Content Evaluation of Compressed Summaries Against Human Annotations Using Dependency Structure Based F-Measure

Technique	F-Measure
Annotator 2	0.829
Annotator 1	0.819
Annotator 4	0.817
Annotator 3	0.808
Annotator 5	0.806
Syntactic with Relevancy	0.759
Keyword Based	0.748
Machine Learning Based	0.707
Relevancy-Driven	0.706
Syntax-Driven	0.664
Baseline: Random Compression 10.5%	0.661
Baseline: Random Compression 21.5%	0.514
Baseline: Random Compression 23.4%	0.514
Baseline: Random Compression 29.8%	0.400
Baseline: Random Compression 43.1%	0.278

all the automatic sentence pruning techniques have performed significantly better than the random word removal baselines (see Table 3). Finally, the syntax complemented with relevancy while removing 23.4% of the words (see Table 1) outperformed all other automatic pruning techniques. This seems to show that it is useful to have predetermined classes of syntactic structures to remove, but they cannot be removed systematically without first verifying their content. The keyword based technique is comparable to the syntax with relevancy in terms of grammatical F-measure but only removed 10.5% of the words (see Table 1).

3 Discussion

The results of Section 3.4 clearly show that a mixture of syntax and relevancy give the best grammatical F-measure given its compression rate. In order to investigate the precision of this approach further, we measured which types of words and phrase structures the annotators removed. Table 4 shows the syntactic structures removed by the annotators along with the compression rate achieved by removing only this structure and the percentage of such structures removed. For example, by removing only prepositional phrases (PPs), the annotators achieve a compression rate of 34.7% on average; but they only removed 12.4% of all prepositional phrases (87.6% of all PPs were left in the compressed summaries). Apart from individual words, noun and verb phrases, the human annotators removed the same syntactic structures as the syntactic sentence pruning techniques, but with a much more subtle selection. In other words, although the removal of PPs account for a great reduction of words (34.7%), only 12.4% of the PPs were actually removed. In effect, this is what relevance score of the

Table 4. Sentence Compression Rates of Different Techniques

Syntactic Structure	Compression %	Removed %
Adverbial Phrases	2.5%	22.0%
Individual Words	2.7%	0.6%
Verb Words	2.8%	1.7%
Adjective Phrases	4.0%	9.4%
Conjoined Clauses	6.4%	9.4%
Appositive Phrases	14.7%	35.2%
Noun Phrases	15.6%	4.1%
Relative Clauses	17.0%	17.2%
Prepositional Phrases	34.7%	12.4%

syntax with relevancy based approach attempted to do; be more subtle about which structure to remove.

It is interesting to note that none of the techniques we have evaluated remove noun and verb phrases. However, as shown in Table 4, humans do remove some. We therefore analyzed these cases to see why they were removed.

In human compressed summaries, a lower percentage of compression was achieved by removing verb phrases compared to noun phrases (2.8% as opposed to 15.6%). After analyzing the noun phrases that were removed, we noted that human annotators seem to remove proper and compound nouns based on their knowledge level. This seems to be subjective for each individual and reflects the annotator's knowledge and perception of the world. As an example, consider the following sentence, pruned by three different annotators:

(4) Annotator 1: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*

(5) Annotator 2: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*

(6) Annotator 3: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*

Here, Annotator 1 has only removed the adjectival phrase *pro-democracy*; while, Annotator 2 has gone a bit further and removed *pro-democracy leader*. Finally, Annotator 3 attempted to remove the entire phrase *pro-democracy leader Aung San* leaving the remaining phrase, *Suu Kyi*. This choice seems to be rather subjective and more influenced by the individuals and is difficult to capture through syntactic pruning rules or relevancy measures or even learned by classifiers.

4 Conclusion and Future Work

In this paper, we have described the evaluation of various sentence pruning approaches and compared them against human compressed summaries. For this task, we used a set of 25 summaries with a word limit of 250, created from the best performing system [28] based on ROUGE scores in the DUC 2007 summarization track. We have used five sets of human compressed summaries, created using the DUC 2007 summaries to evaluate various sentence compression techniques.

First we have evaluated the compression rate of each technique and compared the results against human sentence compression rates. Human sentence compression had an average compression rate of 21.5% which was similar to the compression rate of the syntax with relevancy based technique and machine learning based technique.

We have performed content evaluations using two metrics: ROUGE [23] and a dependency grammar structure based F-measure [24]. The content evaluation using ROUGE showed that even human compressed summaries tend to lose content and the higher the compression rate, the greater the decrease in content compared to the original summaries. This was clearly visible with the baseline systems where we used different compression rates (10.5%, 21.5%, 23.4%, 29.8% and 43.1%). Further, it also showed the weakness of the word-based n-gram ROUGE measure to capture and evaluate attributes such as grammaticality and relevancy of content when it comes automatic summarization.

In our second series of content evaluation, we calculated an F-measure metric based on dependency grammar structures, introduced by [24, 25, 15]. The results were interesting as they showed that this grammar based metric could discriminate the loss of grammaticality of the naïve random removal baselines. The overall results showed that the highest F-measure was not surprisingly achieved by the human annotators with an F-measure of 0.81 and out of all automatic techniques, the syntax with relevancy based sentence compression technique showed the best result with an F-measure of 0.760. Considering that this technique has a similar compression rate to humans and obtained the best grammar structure F-measure, we conclude that the syntax with relevancy based pruning technique seems to model better what humans do. This seems to show that it is useful to have predetermined classes of syntactic structures to remove, but they cannot be removed systematically without first verifying their content.

By analyzing the human compressed summaries, we found that annotators tend to remove syntactic structures more than removing individual words. In addition, these syntactic structures are similar to the syntactic structures typically used in automatic approaches. However, annotators do not remove these syntactic structures systematically but only in certain circumstances. As future work, it would be interesting to further the investigation of which specific structures humans remove and which are kept. Finding discriminating features, other than syntax or relevancy, would be worth looking into.

Acknowledgement. The authors would like to thank the anonymous reviewers for their comments on an earlier version of the paper. This work was financially supported by an NSERC grant.

References

1. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and Methods for Text Simplification. In: Proceedings of COLING 1996, Copenhagen, pp. 1041–1044 (1996)
2. Dorr, B., Zajic, D., Schwartz, R.: Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In: Proceedings of the HLT-NAACL Workshop on Text Summarization, pp. 1–8 (2003)
3. Knight, K., Marcu, D.: Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence 139(1), 91–107 (2002)
4. Gagnon, M., Da Sylva, L.: Text Compression by Syntactic Pruning. In: Lamontagne, L., Marchand, M. (eds.) Canadian AI 2006. LNCS (LNAI), vol. 4013, pp. 312–323. Springer, Heidelberg (2006)
5. Zajic, D., Dorr, B.J., Lin, J., Schwartz, R.: Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. Information Processing and Management 43, 1549–1570 (2007)
6. Harman, D., Liberman, M.: TIPSTER Complete. Linguistic Data Consortium (LDC), Philadelphia (1993)
7. Jaoua, M., Jaoua, F., Belguith, L.H., Hamadou, A.B.: Évaluation de l'impact de l'intégration des étapes de filtrage et de compression dans le processus d'automatisation du résumé. In: Résumé Automatique de Documents. Document numérique, Lavoisier, vol. 15, pp. 67–90 (2012)
8. Jing, H., McKeown, K.R.: Cut and Paste Based Text Summarization. In: Proceedings of NAACL 2000, Seattle, pp. 178–185 (2000)
9. Jing, H.: Sentence Reduction for Automatic Text Summarization. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, Seattle, pp. 310–315 (April 2000)
10. Nguyen, M.L., Phan, X.H., Horiguchi, S., Shimazu, A.: A New Sentence Reduction Technique Based on a Decision Tree Model. International Journal on Artificial Intelligence Tools 16(1), 129–138 (2007)
11. McClosky, D., Charniak, E., Johnson, M.: Effective Self-Training for Parsing. In: Proceedings of HLT-NAACL 2006, New York, pp. 152–159 (2006)
12. Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press (May 1998)
13. Le Nguyen, M., Shimazu, A., Horiguchi, S., Ho, B.T., Fukushi, M.: Probabilistic Sentence Reduction Using Support Vector Machines. In: Proceedings of COLING 2004, Geneva, pp. 743–749 (August 2004)
14. Clarke, J., Lapata, M.: Global Inference for Sentence Compression an Integer Linear Programming Approach. Journal of Artificial Intelligence Research (JAIR) 31(1), 399–429 (2008)
15. Filippova, K., Strube, M.: Dependency Tree Based Sentence Compression. In: Proceedings of the Fifth International Natural Language Generation Conference, INLG 2008, Stroudsburg, PA, USA, pp. 25–32 (2008)
16. Galanis, D., Androutsopoulos, I.: An Extractive Supervised Two-Stage Method for Sentence Compression. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 2010, Los Angeles, California, pp. 885–893 (2010)
17. Perera, P., Kosseim, L.: Evaluating Syntactic Sentence Compression for Text Summarisation. In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, vol. 7934, pp. 126–139. Springer, Heidelberg (2013)

18. Conroy, J.M., Schlesinger, J.D., O'Leary, D.P., Goldstein, J.: Back to Basics: CLASSY 2006. In: Proceedings of the HLT-NAACL 2006 Document Understanding Workshop, New York City (2006)
19. Schlesinger, J.D., O'Leary, D.P., Conroy, J.M.: Arabic/English Multi-document Summarization with CLASSY—The Past and the Future. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 568–581. Springer, Heidelberg (2008)
20. Dunlavy, D.M., Conroy, J.M., Schlesinger, J.D., Goodman, S.A., Okurowski, M.E., O'Leary, D.P., van Halteren, H.: Performance of a Three-Stage System for Multi-Document Summarization. In: Proceedings of the HLT-NAACL 2003 Document Understanding Workshop, Edmonton, Canada, pp. 153–159 (2003)
21. Dang, H.T.: DUC 2005: Evaluation of Question-focused Summarization Systems. In: Proceedings of the Workshop on Task-Focused Summarization and Question Answering, Sydney, pp. 48–55 (2006)
22. Dang, H.T.: Overview of DUC 2006. In: Proceedings of the HLT-NAACL 2006 Document Understanding Workshop (2006)
23. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Moens, M.F., Szpakowicz, S. (eds.) Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop, Barcelona, Spain, pp. 74–81 (July 2004)
24. Riezler, S., King, T.H., Crouch, R., Zaenen, A.: 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, NAACL 2003, Edmonton, Canada, vol. 1, pp. 118–125 (2003)
25. Clarke, J., Lapata, M.: 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, ACL-44, Sydney, Australia, pp. 377–384 (2006)
26. Marneffe, M.C.D., Manning, C.D.: The Stanford Typed Dependencies Representation. In: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, CrossParser 2008, Manchester, pp. 1–8 (2008)
27. Copeck, T., Inkpen, D., Kazantseva, A., Kennedy, A., Kipp, D., Szpakowicz, S.: Catch What You Can. In: Proceedings of Document Understanding Conference (DUC 2007), Rochester, New York, USA (2007)
28. Pingali, P.: K, R., Varma, V.: IIIT Hyderabad at DUC 2007. In: Proceedings of the HLT-NAACL 2007 Document Understanding Workshop, Rochester, New York (2007)