Saarland University

Department of Computational Linguistics

Seminar: **Recent Developments in Computational Discourse Processing**

Compression, Simplification, Fusion:

An Overview of Text Compression Related Tasks

Author:
Patrick Carroll
Matriculation: 2548790

Supervisors:
Dr. Alexis Palmer
Annemarie Friedrich

October 2, 2014

Abstract

The abstract will probably have to be filled out at the very end of writing the paper, because I am not sure what shape it will take until I have some things written down on the page. Hopefully this is not an issue.

The goal of this paper is to provide an overview of three major text reduction techniques that are currently being developed in the Natural Language Processing world.

Contents

1 Introduction

In the Natural Language Processing community there is currently an open frontier of research tasks focused on ways in which texts can be compressed or transformed from their original versions into some form of distilled output. The general goal of these tasks is to reduce a single text or several texts in such a way that important information is preserved but some aspect of complexity is reduced. Over the course of this paper, I intend to give an overview of three of these tasks: text compression, text simplification, and sentence fusion. While exploring these tasks, I also intend to focus on the use of discourse level information, which, when applicable, represents a promising means of identifying important information in these compression related tasks.

Let me begin by providing some basic definitions of what each task entails, and pointing out which publications this paper will focus on. I'll begin with text compression as it is the most straightforward definition and will serve as a reference point for how the other two tasks diverge. Text compression was originally defined as a sentence level operation intended to reduce the length of a sentence and/or produce a sentence summary which preserves the most important information while also remaining grammatical (Jing, 2000). In order to provide more flexible compression of whole documents, recent systems have been developed which used discourse information across the whole document, rather than just on the sentence level, in order to make more informed choices about what words get compressed. In this paper, I will be focusing my attention one such example by Clarke and Lapata (2010).

The task of text simplification can be defined as: given a source sentence we aim to produce a simplified version of that sentence with simpler vocabulary and sentence structure while preserving the main ideas in the original sentenceFeng (2008). In contrast to text compression, the procedure for simplification is not as clear cut as just deleting/retaining words (although deletion may play some role). Often the task calls for splitting long sentences into a series of smaller sentences, replacing semantically difficult or ambiguous words with simpler ones, or re-phrasing a sentence to change it's syntactic structure (Coster and Kauchak, 2011). For this paper I will focus on 2 examples. The first example from from Siddharthan (2006) uses discourse information to break down syntactically complex sentences into smaller sub-sentences. The second example from Coster and Kauchak (2011) attempts to reduce the semantic complexity of a text, and frames the simplification task as a translation from standard to simple English using parallel corpora taken from Wikipedia.[1]

The task of sentence fusion may be seen as somewhat similar to text compression, but applied over a multi-document collection of texts. However, to adequately describe the

[1] http://simple.wikipedia.org, http://en.wikipedia.org

task of sentence fusion, it is best to divide it into a two step process. The first step is a preprocessing step of paraphrase detection, and the second step is the sentence fusion itself. In paraphrase detection, the aim is to search a multi-document database and identify clusters of equivalent sentences or sentences fragments (Regneri and Wang, 2012). In the sentence fusion step, the aim is to create a summary sentence for each paraphrase cluster which maximizes common information across documents, while minimizing redundancy. This is done by fusing parts of sentences from the cluster together to make one concise, cohesive sentence.(Filippova and Strube, 2008). Using the two articles cited above as my examples, I will demonstrate how both steps in the sentence fusion process can benefit from the use of discourse information.

Continuing on from these basic definitions, the remaining sub-sections of the introduction will look in greater detail at the real world applications, the models which are used to tackle each task, and the data for training and testing. Following the introduction, section 2 will detail the systems from each of the publication mentioned above, and how they use discourse information (if applicable). Section 3 is reserved for discussion of the relations between compression based tasks, conclusion on the current state of these technologies, and speculation on future work.

1.1 Applications

As mentioned in the previous section, all three tasks share an over-arching purpose of transforming a text from it's original content into a condensed form. Just what that condensed text can be useful for is the purpose of this sub-section. Here we will look at what applications the tasks can be used for, and how they may improve that process.

A good starting point when talking about applications is to look at how text summarization can be augmented by both sentence compression and sentence fusion. As a baseline, most text summarization systems use extractive approaches which shorten a document by retaining important sentences, and discarding everything else. While extractive approaches guarantee grammatical output on the sentence level, it's often not a fine-grained enough solution for some summarization goals. For instance, if the goal is to have a summary with maximum compression of text length, the simple extractive approach falls short. This is because some texts will inevitably contain sentences with a mix of important and superfluous information. Sentence compression was envisioned as a means to tackle these cases. By compressing the sentences which are output from extractive summarization systems, a sentence compression system can further reduce summary length for more concise text summaries(Jing, 2000). This is in effect a double compression pipeline which first cuts out un-necessary sentences using extraction, then cuts out un-necessary words using compression.

When the goal of a text summarization system is scaled up to summarize a collection of documents,the baseline extractive approach runs into a similar problem of balancing important information with summarization brevity. In the baseline multi-document extractive approach, sentences from across all documents are clustered based on their similarity to each other. The summary is created by picking sentences from the clusters which are the best representations of important information. Like in the previous example, sentences are selected whole, which often results is an adequate summary with some superfluous text, or a summary which sacrifices important information in favor of brevity (Filippova and Strube, 2008). In these cases sentence fusion was envisioned as a means to pick the best snippets of informative text from a cluster and fuse them into new sentence, thus greatly reducing the amount of redundant or unneeded text.**?**.

While text summarization plays a large role in the development of both compression and sentence fusion, there are many other applications which also call for the use of compression based tasks. Transforming texts for greater readability has also been a popular topic of research, and has made good use of text compression and text simplification. In the case of text simplification, the goal of improving readability is often focused on people with reading comprehension difficulties such as children and foreign language learners, or those with cognitive impairments such as aphasics(Feng, 2008). In such applications, the purpose of text simplification is to reduce the semantic complexity of a text, and/or to alleviate some memory load issues by breaking complex sentences down into shorter, more easily processed sentences. Some other readability applications include using text compression to reducing text length for display on PDA's (**?**) or as a reading aid for the blind (**?**). In these cases, compression will suffice because it's readability issues arise from text length rather than text difficulty.

1.2 Models

This subsection is intended to describe how the problem of each task is modeled. By looking at what operations are to be performed, and the constraints governing those operations, one can see how the three tasks are related.

Beginning with compression, the task is typically viewed as a word deletion operation. According to Knight and Marcu (2002), given an input sentence of words $x = x_1, x_2...x_n$ the aim is to produce a compression which is some sub-set of these words. The constraints in this model can be viewed conceptually as the rules which decide which words are kept, and which are deleted.

The approach used to solve the problem of text simplification expands upon the deletion problem mentioned in the previous paragraph. In addition to the deletion operation, the operations of reduction, rewording, reordering, and insertion may also be included in modeling the task(Siddharthan, 2006)(Coster and Kauchak, 2011). In terms of constraints which determine what operation to perform, the example system from Siddharthan (2006) uses hand coded rules based on syntatic information, while the system from Coster and Kauchak (2011) is a data driven approach which learns rules for the operations from parallel corpora or standard and simple texts.

The approaches used for sentence fusion must be viewed as two sub tasks. The task of paraphrase detection uses a

1.3 Data

Highlight the kind of data used to train and/or model the problem for each text reduction task. Is the data directly used to train a system, or is it simply used as a frame of reference for un-supervised learning. What kind of data is used for validation and evaluating the systems?

This section will focus on the data used to train

1.4 Natbib citations

Within a text, you can say that Lin and Pantel (2001) found out something. Or you can just state the thing, and then put the author in parentheses (see Szpektor et al., 2004).

2 Compression Related Tasks

2.1 Compression

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

2.2 Simplification

2.3 Paraphrasing/Fusion

Figure 1. The saarland uni logo.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

3 Discussion

3.1 Conclusion

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

3.2 Thoughts on Future Work

References

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, 31(1):399–429.

Clarke, J. and Lapata, M. (2010). Discourse constraints for document compression. *Comput. Linguist.*, 36(3):411–441.

Coster, W. and Kauchak, D. (2011). Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Feng, L. (2008). Text simplication: A survey. Technical report, CUNY.

Filippova, K. and Strube, M. (2008). Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 177–185, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 310–315, Stroudsburg, PA, USA. Association for Computational Linguistics.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.

Lin, D. and Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 323–328, San Francisco, CA.

Regneri, M. and Wang, R. (2012). Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 916–927, Stroudsburg, PA, USA. Association for Computational Linguistics.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004). Scaling web-based acquisition of entailment relations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 41–48, Barcelona, Spain.