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 multi-document text paraphrasing / sentence fusion. While exploring these tasks, I also intend to focus on the use of discourse analysis, which, when applicable, represents a promising means of identifying important information in these compression related tasks.

Let us then begin with some basic definitions of what each task entails. Text compression was originally defined as a sentence level operation with the goal of producing a sentence summary which preserves the most important information and also remains grammatical **?**. This idea of sentence compression has been expanded by researchers in the intervening years to also include information beyond the sentence **?** by using discourse constraints to have document level information to inform sentnece compression.

as a reduced text is one in which the total number of words is reduced, while still preserving the important information and retaining grammaticality. In the task of text simplification, a reduced text would be semantically and/or syntactically less complex, and may also be reduced in length (though this is not a necessary condition). In the task of paraphrasing, a collection of texts is searched for equivalent sentences which represent important information. From thes sentences an abstractive summary can be generated by fusing pieces of these sentences together. Over the course of this paper I will highlight the commonalities and differences among these tasks, and focus special attention on how discourse information has been used in each task.

single text or multiple texts can be compressed, simplified or fused together. This field of research can be seen, in a general sense, as ways in which a source text (or collection of texts) can be transformed into a less complex, more managable output text designed to meet some goal. For the sake of consistience I will refer to these tasks as Because of the diverse end goals a researcher may be seeking by reducing a text's complexity, the methods, data representations,

:

B̂ecause of the diverse end goals a researcher may be aiming for, each of the three tasks (compression, simplification, paraphrasing) approach the goal of transforming the source text in a different manner.

1.1 Goals

As mentioned in the previous section, all three tasks being discussed share an over-arching purpose of reducing a text from it's original content into a more manageable and useful sequence of words. This reduced sequence of words (which I will refer to subsequently as a *reduced text*) should naturally vary in it's characteristics depending on what the end goal of the application may be. Thus we are interested in what applications the different tasks (compression, simplification, paraphrasing/fusion) have been developed to handle at present, and what future applications may also be aided by these tasks.

In the case of text compression, current applications include compressing texts in tandem with text summarization systems to improve conciceness **?**, or as a reading aid for the blind **?**. The task of text simplification has been employed to reduce a text's reading difficulty level for children**?**.or those with reading imparements **?**. It has also been used for improving information retrieval tasks, including medical document information retrieval **?**. In the case of mulit-document text paraphrasing, systems are designed to reduce text from multiple sources in order to create an abstractive summary.**?**.

1.2 Approaches

Talk about the way that the problems are modeled for each category of text reduction. Then talk about some of the algorithms (and possibly machine learning paradigms) used to solve the tasks of compression, simplification, and paraphrasing. When there is overlap point it out, and also make note of when there are drastically diverging.

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?

1.4 Use of Discourse Information

Go into depth about what systems make use of Discourse level informaiton, either directly in processing of the text, or perhaps in a more limited aspect in the evaluation of the output. Also mention versions of text reduction that do not make use of any Discourse information. Are they any better? Is discourse information at this point not terribly helpful to solving the task?

1.5 Natbib citations

Within a text, you can say that **?** found out something. Or you can just state the thing, and then put the author in parentheses (see **?**).

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