Saarland University

Department of Computational Linguistics

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

Compression, Simplification, Paraphrasing:

An Overview of Automatic Text Reduction

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

The term text reduction is intended as a broad category which encompasses three distinct tasks: text compression, text simplification, and multi-document text paraphrasing. The three tasks all share a common goal of reducing a text from it's original form into a more manageable and useful format suited to a specific need. In the case of text compression, texts are reduced in order to improve reading times,**?** or to adapt text to smaller devices**?**. For Text simplification, the goal may be to reduce a text's complexity for children**?**.or those with reading imparements **?**. It has also been recently used for improving 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.**?**.

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. In the task of text compression, 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.

1.1 Goals

Here talk about the Goals of each category, and how that has come about over the course of research on the subject

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 Fusce mauris

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.

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.