575 Report D2

Rachel Hantz, Yi-Chien Lin, Yian Wang, Tashi Tsering, Chenxi Li

Department of Linguistics
University of Washington
Seattle, WA USA
{hantz1rk,yichlin,wangyian,tashi0,cl91}@uw.edu

Abstract

A short high-level overview of the paper, usually 150 words or so.

1 Introduction

Briefly overview the questions you are approaching, summarize the main conclusions, and give an overview of the paper.

This is an example of citation. ... (Paul et al., 2010).

2 System Overview

A description of the major design, methodological, and algorithmic decisions in your project. It often includes a schematic of the system architecture.

3 Approach

This section should provide the details of the major subcomponents of your system.

3.1 Preprocessing

In order to use the articles that we will summarize, we first needed to pre-process them. We accomplished this with a script that implemented two overall steps: process and tokenization.

Processing takes in the path of input xml file and extracts the document ID. Specifically, we imported xml.dom.minidom to parse the path of input xml file: we called getElementsByTagName() to obtain the elements under docsetA and called getAttribute() to obtain each document ID.

Tokenization takes in a document ID and output a file of desired format needed for our later tasks. After locating the xml document in corpora, we used xml.etree.ElementTree to create a tree for the xml document. On the root node, we obtained text content of a part by matching the tag name (e.g., node.tag == HEADLINE). For tokenization, we just used nltk.sent_tokenize() to break the paragraph and and nltk.word_tokenize() to break each sentence.

The tokenization schema produces output files for each article. Each file starts with a tokenized headline and has a single tokenized sentence per line. Paragraphs are separated with a single blank line.

4 Results

Results of the formal evaluation of your system and components.

5 Discussions

Error analysis and assessment of the strengths and weaknesses of the different components.

6 Conclusion

Summarize the main points and look ahead. What would your next steps be?

References

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 66–76.

Appendix A

For D1: Rachel went through the tutorial of git; Yi-Chien and Yian showed the basics of overleaf; Tashi and Chenxi wrote up the submission pdf.

For D2: Yian did the coding part, with some help from Rachel and Yi-Chien; Yi-Chien posted tutorial on setting up Anaconda environment on Patas for the group; Tashi made the slides for presentation; Chenxi wrote up the report D2. Rachel will be presenting in class.

Appendix B

Link to the code repository on github: https://github.com/rhantz/575_

summarization
Off-the-shell tools used in code:

- xml to parse the path of input file and the xml document in corpora
- nltk for sentence and word tokenization
- os for system operation on Patas

Appendix C

Problem 1:
Description: documents in AQUAINT corpora are not rooted, causing parsing to fail.
Solution: created a root node by inserteing <tag> at the start and appending <\tag> in the end.

Problem 2:
Description: code fails to run on Patas for some group members
Solution: set Anaconda on Patas to ensure people have the same environment.