Toward Automatic Classification of Metadiscourse

Rui Correia[1,2,4], Nuno Mamede[1,2], Jorge Baptista[2,3], and Maxine Eskenazi[4]

[1] Universidade de Lisboa – Técnico Lisboa, Portugal
[2] Laboratório de Língua Falada – INESC-ID Lisboa, Portugal
[3] Universidade do Algarve, Portugal
[4] Language Technologies Institute, Carnegie Mellon University, USA

Introduction

What is Metadiscourse?

- linguistic material used to help the audience organize, interpret and evaluate the information given [Crismore et al. 1993]
- occurs in both written and spoken language

Written	Spoken
In this paper we present…	*I'm going to talk about…*
It is important to note that…	*The take home message is…*
In sum…	*I'd like to conclude this talk with…*

Goal

- focus on spoken language
- develop a unified metadiscourse classification platform
  - differentiate strategies according to their function in discourse

Motivation

- teaching how to make effective presentations
- aid to discourse analysis, or summarization tasks

Taxonomy and Corpus

Ädel's (2010) taxonomy of metadiscourse

- functional organization
  - metadiscourse as discourse functions
- 23 categories
- built based on academic papers and university lectures
- in this work:

Introducing Topic; Concluding topic; Exemplifying; Emphasizing

TED talks

- well prepared presentations
- short and self-contained
- target broad audience
- multilingual
- 730 talks transcribed in English (subtitled by the TED community)

Crowdsourced Annotation

Amazon Mechanical Turk (AMT)

- 4 different tasks uploaded (one per category)
  - lessen workers' cognitive load
- talks divided into segments of 300 words
- workers asked to click on words representative of each category

Quality Control

- training sessions (with feedback)
  - filter out workers with unsuccessful results
- gold standard
- self-confidence report (5-point Likert scale)
- 3 workers per segment

Annotation Results

Category	workers in agreement ≥ 1	≥ 2	$= 3$	conf	κ
INTRO	1,894	**1,159**	600	3.95	0.64
CONC	1,045	**628**	285	4.00	0.60
EXMPL	1,764	**1,327**	720	3.94	0.72
EMPH	3,450	**2,580**	750	3.99	0.58

κ: inter-annotator agreement
*annotators agree if the intersection of their selected words is not empty

Experimental Setup

- crowd annotations used as training data
  - majority vote in case of disagreement

Features

- n-grams
  - Part-Of-Speech (POS)
  - Word Lemmas
  - Word Tokens

Settings

- Decision Trees
  - C4.5 algorithm as implemented in WEKA
- Stop Words not discarded
- Feature Reduction
  - Information Gain > 0.0025

Results

- 10-fold cross-validation
- balanced set of positive and negative examples
- report accuracy

Category	POS *n-grams* 1	2	3	Lemma *n-grams* 1	2	3	Token *n-grams* 1	2	3
INTRO	79.6	85.1	86.4	91.6	92.6	92.2	92.3	**92.7**	**92.7**
CONC	65.8	66.4	68.1	84.8	**86.9**	86.1	84.9	86.7	**86.9**
EXMPL	61.8	65.4	68.2	94.1	94.3	94.3	94.2	94.7	**94.9**
EMPH	67.3	68.2	68.9	77.9	79.6	**79.8**	79.4	79.5	**79.8**

Feature Combining

- combination of previous settings
- statistically significant improvement ($p < 0.05$) for Emphasizing
  - Lemma and Token 3-gram model with 81.5%

Highest ranked n-grams

- Introducing Topic
  - *I*, *to*, *about*, *talk*, *show you*, *tell you*, and *going to*
- Concluding topic
  - *finally*, *end*, *last*, *conclude*, *so in*, *so to*, and *my time*.
- Exemplifying
  - *example*, *imagine*, *instance*, *suppose*, *look at*, *give you*, *such as*, *if you were*, and *think about the*
- Emphasizing
  - *important*, *emphasize*, *to focus*, *point out*, and *idea is*

Conclusions

- lexical information proved to be representative of metadiscourse
- use of some of the markers is strongly conventionalized
  - decision tree with 19 rules for Exemplifying *vs.* 200 rules for Emphasizing
- accuracy matches the quality of human identification observed

Future Work

- additional metadiscursive categories
- improve classification
  - dependencies
  - dictionaries of discourse clues
- explore metadiscourse in European Portuguese

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and by FCT project CMU-PT/HuMach/0053/2008