

# Discourse Processing

# Synthesis Lectures on Human Language Technologies

## Editor

**Graeme Hirst**, *University of Toronto*

Synthesis Lectures on Human Languages Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

## Discourse Processing

Manfred Stede

2011

## Bitext Alignment

Jörg Tiedemann

2011

## Linguistic Structure Prediction

Noah A. Smith

2011

## Learning to Rank for Information Retrieval and Natural Language Processing

Hang Li

2011

## Computational Modeling of Human Language Acquisition

Afra Alishahi

2010

## Introduction to Arabic Natural Language Processing

Nizar Y. Habash

2010

### Cross-Language Information Retrieval

Jian-Yun Nie

2010

### Automated Grammatical Error Detection for Language Learners

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault

2010

### Data-Intensive Text Processing with MapReduce

Jimmy Lin and Chris Dyer

2010

### Semantic Role Labeling

Martha Palmer, Daniel Gildea, and Nianwen Xue

2010

### Spoken Dialogue Systems

Kristiina Jokinen and Michael McTear

2009

### Introduction to Chinese Natural Language Processing

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang

2009

### Introduction to Linguistic Annotation and Text Analytics

Graham Wilcock

2009

### Dependency Parsing

Sandra Kübler, Ryan McDonald, and Joakim Nivre

2009

### Statistical Language Models for Information Retrieval

ChengXiang Zhai

2008

© Springer Nature Switzerland AG 2022  
Reprint of original edition © Morgan & Claypool 2012

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

DiscourseProcessing  
Manfred Stede

ISBN: 978-3-031-01016-3 paperback  
ISBN: 978-3-031-02144-2 ebook

DOI10.1007/978-3-031-02144-2

A Publication in the Springer Nature series  
*SYNTHESIS LECTURES ON ADVANCES IN AUTOMOTIVE TECHNOLOGY*

Lecture#15  
SeriesEditor:GraemeHirst,*UniversityofToronto*  
SeriesISSN  
SynthesisLecturesonHumanLanguageTechnologies  
Print1947-4040 Electronic1947-4059

# Discourse Processing

Manfred Stede  
University of Potsdam

*SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #15*

## ABSTRACT

*Discourse Processing* here is framed as marking up a text with structural descriptions on several levels, which can serve to support many language-processing or text-mining tasks. We first explore some ways of assigning structure on the document level: the logical document structure as determined by the layout of the text, its genre-specific content structure, and its breakdown into topical segments. Then the focus moves to phenomena of local coherence. We introduce the problem of coreference and look at methods for building chains of coreferring entities in the text. Next, the notion of coherence relation is introduced as the second important factor of local coherence. We study the role of connectives and other means of signaling such relations in text, and then return to the level of larger textual units, where tree or graph structures can be ascribed by recursively assigning coherence relations. – Taken together, these descriptions can inform text summarization, information extraction, discourse-aware sentiment analysis, question answering, and the like.

## KEYWORDS

text structure, document structure, topic segmentation, coreference, anaphora resolution, coherence relation, discourse parsing

# Contents

	<b>Acknowledgments</b> .....	<b>ix</b>
<b>1</b>	<b>Introduction</b> .....	<b>1</b>
<b>2</b>	<b>Large Discourse Units and Topics</b> .....	<b>7</b>
2.1	Genre-induced text structure .....	7
2.1.1	Logical document structure .....	9
2.1.2	Content zones .....	13
2.1.3	Example: Scientific papers .....	13
2.1.4	Example: Film reviews .....	15
2.2	Topic-based segmentation .....	16
2.2.1	Introduction: “What is this all about?” .....	16
2.2.2	Exploiting surface cues .....	19
2.2.3	Lexical chains .....	22
2.2.4	Word distributions .....	28
2.2.5	Probabilistic models of segmentation and topics .....	34
2.2.6	Combining evidence .....	36
2.3	Summary .....	37
<b>3</b>	<b>Coreference Resolution</b> .....	<b>39</b>
3.1	Reference and coreference: An overview .....	40
3.2	Corpus annotation .....	46
3.3	Entity-based local coherence .....	51
3.4	Determining anaphoricity and familiarity status .....	54
3.5	Rule-based methods for resolving nominal anaphora .....	57
3.5.1	Matching proper names .....	57
3.5.2	Pronoun resolution .....	58
3.5.3	Resolving definite noun phrases .....	61
3.5.4	Web-assisted resolution of ‘other’-anaphora .....	63
3.6	Statistical approaches to coreference resolution .....	65
3.6.1	Features .....	66
3.6.2	Mention-pair models .....	68

	3.6.3 Alternative models .....	71
3.7	Evaluation .....	74
3.8	Summary .....	75
<b>4</b>	<b>Small Discourse Units and Coherence Relations .....</b>	<b>79</b>
4.1	Coherence relations .....	79
4.2	Segmentation: Finding elementary discourse units .....	87
	4.2.1 Defining EDUs .....	87
	4.2.2 A subproblem: Attribution .....	90
	4.2.3 Automatic EDU segmentation .....	92
4.3	Recognizing coherence relations .....	97
	4.3.1 Connectives: An introduction .....	97
	4.3.2 Identifying connectives .....	101
	4.3.3 Interpreting connectives .....	103
	4.3.4 Detecting implicit coherence relations .....	110
	4.3.5 Finding relations: The problem at large .....	112
4.4	Coherence-relational text structure .....	113
	4.4.1 Trees .....	114
	4.4.2 Parsing coherence-relational trees .....	117
	4.4.3 Graphs .....	123
4.5	Summary: Guessing or underspecifying? .....	126
<b>5</b>	<b>Summary: Text Structure on Multiple Interacting Levels .....</b>	<b>129</b>
<b>A</b>	<b>Sample text .....</b>	<b>133</b>
	<b>Bibliography .....</b>	<b>137</b>
	<b>Author's Biography .....</b>	<b>155</b>

# Acknowledgments

The author is grateful to the anonymous reviewers and to the series editor for their thoughtful and constructive suggestions to improve an earlier version of the manuscript. Also, thanks are due to several readers who provided comments on earlier versions of individual sections: Heike Bieler, Markus Egg, Thomas Hanneforth, Manfred Klenner, Constantin Orasan, Maite Taboada, and Sebastian Varges.

Manfred Stede  
November 2011