2017

From Discourse Structure To Text Specificity: Studies Of Coherence Preferences

Junyi Li
University of Pennsylvania, jyjessyli@gmail.com

From Discourse Structure To Text Specificity: Studies Of Coherence Preferences

Abstract
To successfully communicate through text, a writer needs to organize information into an understandable and well-structured discourse for the targeted audience. This involves deciding when to convey general statements, when to elaborate on details, and gauging how much details to convey, i.e., the level of specificity. This thesis explores the automatic prediction of text specificity, and whether the perception of specificity varies across different audiences.

We characterize text specificity from two aspects: the instantiation discourse relation, and the specificity of sentences and words. We identify characteristics of instantiation that signify a change of specificity between sentences. Features derived from these characteristics substantially improve the detection of the relation. Using instantiation sentences as the basis for training, we propose a semi-supervised system to predict sentence specificity with speed and accuracy. Furthermore, we present insights into the effect of underspecified words and phrases on the comprehension of text, and the prediction of such words.

We show distinct preferences in specificity and discourse structure among different audiences. We investigate these distinctions in both cross-lingual and monolingual context. Cross-lingually, we identify discourse factors that significantly impact the quality of text translated from Chinese to English. Notably, a large portion of Chinese sentences are significantly more specific and need to be translated into multiple English sentences. We introduce a system using rich syntactic features to accurately detect such sentences. We also show that simplified text is more general, and that specific sentences are more likely to need simplification. Finally, we present evidence that the perception of sentence specificity differs among male and female readers.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Computer and Information Science

First Advisor
Ani Nenkova

Second Advisor
Mitchell P. Marcus

Keywords
computational linguistics, discourse, natural language processing, specificity

FROM DISCOURSE STRUCTURE TO TEXT SPECIFICITY: STUDIES OF COHERENCE PREFERENCES

Junyi Jessy Li

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Ani Nenkova, Associate Professor, Computer and Information Science

Graduate Group Chairperson

_____

Lyle Ungar, Professor, Computer and Information Science

Dissertation Committee:

Mitchell P. Marcus, Professor, Computer and Information Science (Chair)

Mark Liberman, Professor, Linguistics, Computer and Information Science

Bonnie Webber, Adjunct Professor, Computer and Information Science; Professor, School of Informatics, University of Edinburgh

Marine Carpuat, Assistant Professor, Computer Science, University of Maryland (External)

Jacob Eisenstein, Assistant Professor, School of Interactive Computing, Georgia Institute of Technology (External)

FROM DISCOURSE STRUCTURE TO TEXT SPECIFICITY:
STUDIES OF COHERENCE PREFERENCES

To my family

Acknowledgements

First of all, I would like to thank my advisor Ani Nenkova. I started working with Ani when I was a masters student at Penn. If it were not for her encouragement and support, and the endless fun working with her, I would not even have started my Ph.D. I thank her for her constant dedication and nurturing: her patient teachings and guidance in research, our weekly meetings, rounds of comments for my writing pieces, and advice in serving the research and education community at Penn and beyond. I am forever grateful for her belief in me, and I hope to pass on her passion for research and mentorship to my future students.

I am grateful to Bonnie Webber for her guidance since the early years of my Ph.D. Bonnie has given invaluable advice for many pieces of my then half-baked work and met with me during her visits to Penn. Her mentorship helped to shape my research. I thank Mitch Marcus and Mark Liberman for their steadfast support throughout my years at Penn; discussions with them are always filled with insights and delights. I thank Marine Carpuat and Jacob Eisenstein, whose generous help and valuable feedback improved this thesis.

I would like to thank my collaborators, without whom many projects would not have happened: Sumit Basu, Leila Bateman, Marine Carpuat, Charles Jacobs, Matthew Lease, Iain Marshall, An Thanh Nguyen, Ben Nye, Bridget O'Daniel, Julia Parish-Morris, Amanda Stent, Kapil Thadani, Lucy Vanderwende, Byron Wallace, Yi Wu, Yinfei Yang, and Wenli Zhao.

I thank my mentors during my internships, whose support has made these internships wonderful experiences: Sumit Basu and Lucy Vanderwende at Microsoft Research, Amanda Stent and Kapil Thadani at Yahoo! Labs. I thank Yashar Mahdad, Dragomir Radev and Joel Tetreault for the fun and constructive conversations when I was at Yahoo! Labs, and Mausam for being my mentor at the AAAI'16 Doctoral Consortium.

I am honored to be a part of the outstanding linguistics and computational linguistics community at Penn. I enjoyed many conversations with Aravind Joshi, who has guided and influenced me with his knowledge, wisdom and enthusiasm. I have also received much support from Chris Callison-Burch, Ariani Di Felippo, Florian Schwarz, Muffy Siegel, Lyle Ungar, and Charles Yang. I would like to thank my wonderful fellow students and colleges in the NLP group: Houwei Cao, Spencer Caplan, Anne Cocos, Paramveer Dhillion, Jie Gao, Kai Hong, Jordan Kodner, Constantine Lignos, Xi Victoria Lin, Annie Louis, Ellie Pavlick, Emily Pitler, Daniel Preotiuc, Neville Ryant, Andy Schwartz, Joao Sedoc, and Wei Xu.

Penn has amazing staff who have helped me on numerous occasions with patience, accuracy and efficiency. Special thanks to Mike Felker, Cheryl Hickey and Rita Powell from the department of Computer and Information Science, and Amanda Phipps from the International Student and Scholar Services.

I am grateful to those who have inspired me in my earlier education. Special thanks to Eytan Adar (University of Michigan), for introducing me to research as an undergraduate student; Shensheng Zhang (Shanghai Jiao Tong University), for introducing me to programming; and Wenjian Wang (Shanghai Wei Yu High School), for his patient and constant guidance throughout my high school years.

I would like to thank other colleagues and friends, including Arthur Azevedo De Amorim, Waleed Ammar, Ang Chen, Chen Chen, Sanjian Chen, Loris D'Antoni, Luheng He, Yuening Hu, Jennifer Hui, Charlie Jin, Sarfraz Khurshid, Sungjin Lee, Xiang Li, Daniella Marinov, Darko Marinov, Lara Marinov, Fei Miao, Sasa Misailovic, Salar Moarref, Lu Pan, Ankur Parikh, Jennifer Paykin, Venetia Pliatsika, Robert Rand, Jenn Ruiz, Nick Ruiz, Hongbo Zhang, Yemin Tang, Zhongni Tang, Shan Wang, William Wang, Steven Wu, Meng Xu, Mark Yatskar, Mabel Zhang and Nan Zheng.

Above all, my deepest gratitude goes to my family. I thank *Milos*, my father *Gongsheng* and my mother *Qiufang* for their unwavering love, support and encouragement. I thank *David* for bringing so much wonder and joy into my life. I thank my extended family and Milos' family for their care and support.

ABSTRACT

FROM DISCOURSE STRUCTURE TO TEXT SPECIFICITY:
STUDIES OF COHERENCE PREFERENCES

Junyi Jessy Li

Ani Nenkova

To successfully communicate through text, a writer needs to organize information into an understandable and well-structured discourse for the targeted audience. This involves deciding when to convey general statements, when to elaborate on details, and gauging how much details to convey, i.e., the level of specificity. This thesis explores the automatic prediction of text specificity, and whether the perception of specificity varies across different audiences.

We characterize text specificity from two aspects: the INSTANTIATION discourse relation, and the specificity of sentences and words. We identify characteristics of INSTANTIATION that signify a change of specificity between sentences. Features derived from these characteristics substantially improve the detection of the relation. Using INSTANTIATION sentences as the basis for training, we propose a semi-supervised system to predict sentence specificity with speed and accuracy. Furthermore, we present insights into the effect of underspecified words and phrases on the comprehension of text, and the prediction of such words.

We show distinct preferences in specificity and discourse structure among different audiences. We investigate these distinctions in both cross-lingual and monolingual context. Cross-lingually, we identify discourse factors that significantly impact the quality of text translated from Chinese to English. Notably, a large portion of Chinese sentences are significantly more specific and need to be translated into multiple English sentences. We introduce a system using rich syntactic features to accurately detect such sentences. We also show that simplified text is more general, and that specific sentences are more likely to need simplification. Finally, we present evidence that the perception of sentence specificity differs among male and female readers.

Contents

Acknowledgements **iv**

1 Introduction **1**

1.1 Thesis contributions 3

1.2 Thesis organization 5

2 Notions of specificity in text **8**

2.1 Discourse relations 9

2.2 Information organization 11

2.3 Fine-grained linguistic structure 13

3 From discourse relation to text specificity **15**

3.1 Discourse relations in the Penn Discourse Treebank 17

3.2 Recognizing INSTANTIATION 18

3.2.1 Data and methodology 20

3.2.2 Characteristics of INSTANTIATION 20

3.2.3 Experiments 23

3.2.4 Discussion: textual entailment in INSTANTIATION 25

3.2.5 Conclusion 26

3.3 SPECITELLER: predicting sentence specificity 26

3.3.1 Data 27

3.3.2 Light features 27

3.3.3 Supervised learning results 29

3.3.4 Semi-supervised learning via co-training 30

3.3.5 Semi-supervised learning results . . . 33
3.3.6 Discussion . . . 34
3.3.7 Conclusion . . . 36
3.4 Discussion: INSTANTIATION vs. SPECIFICATION . . . 36

4 Fine-grained text specificity and its connection with discourse 41
4.1 Annotation and corpus analysis . . . 43
4.1.1 Goal . . . 43
4.1.2 Methodology and corpus summary . . . 44
4.1.3 Corpus statistics . . . 45
4.1.4 Discourse analysis of underspecification . . . 48
4.1.5 Conclusion . . . 52
4.2 Predicting subsentential specificity . . . 52
4.2.1 Related work . . . 54
4.2.2 Attention network for token specificity . . . 55
4.2.3 Named entities and pronouns . . . 56
4.2.4 Systems and settings. . . . 57
4.2.5 Sentence specificity prediction. . . . 58
4.2.6 Predicting question tokens . . . 59
4.2.7 Conclusion . . . 64
4.3 Discussion and future work . . . 64

5 Coherence preferences: cross-lingual 66
5.1 Assessing the Discourse Factors that Influence the Quality of Machine Translation . . . 67
5.1.1 Data and experiment settings . . . 68
5.1.2 Sentence length and HTER . . . 69
5.1.3 When a sentence becomes discourse . . . 70
5.1.4 Explicit discourse relations . . . 71
5.1.5 Number of connectives . . . 72
5.1.6 Relation senses . . . 73

5.1.7 Human edits of discourse connectives . . . 73
5.1.8 Discussion and conclusion . . . 75
5.2 Discourse vs. sentence: identifying content-heavy sentences . . . 76
5.2.1 Data . . . 78
5.2.2 Content-heavy sentences: definition . . . 79
5.2.3 A challenge for MT . . . 80
5.2.4 Comma usage and heavy sentences . . . 81
5.2.5 Features to characterize content-heavy sentences . . . 84
5.2.6 Recognizing content-heavy sentences . . . 86
5.2.7 A challenge for MT: revisited . . . 88
5.2.8 Conclusion . . . 89
5.3 Discussion and future work . . . 89

6 The organization of specific information 91
6.1 Data and settings . . . 92
6.2 Specificity . . . 93
6.2.1 Consistency across reference translations . . . 93
6.2.2 Heavy vs. non-heavy sentences . . . 95
6.2.3 Intra-sentential specificity . . . 96
6.3 Instantiation and other discourse relations . . . 96
6.3.1 Implicit Instantiation in Chinese . . . 97
6.3.2 Content-heavy sentences and Instantiation arguments . . . 98
6.3.3 Relations across split components of heavy sentences . . . 99
6.4 Predicting heaviness . . . 100
6.5 Conclusion . . . 102

7 Coherence preferences: monolingual 103
7.1 Sentence specificity and text simplification . . . 104
7.1.1 Specificity as simplification objective . . . 104
7.1.2 Identifying simplification targets . . . 106
7.1.3 Conclusion . . . 107

7.2 Gender, Autism Quotient scores and the perception and production of text specificity . . . 107
7.2.1 Overview . . . 108
7.2.2 Subjects . . . 109
7.2.3 Specificity perception . . . 110
7.2.4 Specificity of text produced . . . 113
7.2.5 Discussion and conclusion . . . 114
7.3 Discussion and future work . . . 115

8 Conclusion **116**
8.1 Summary of contributions . . . 116
8.2 Future work . . . 118

References **121**

List of Tables

2.1 Definition of ELABORATION in Rhetorical Structure Theory (Mann and Thompson, 1988). 9

2.2 Definition of INSTANTIATION and SPECIFICATION in the Penn Discourse Treebank (Miltsakaki et al., 2008). 9

3.1 Average numbers of words, percentages of rare words and percentages of gradable adjectives per sentence for: first sentences of INSTANTIATION, second sentences of INSTANTIATION, and non-INSTANTIATION sentences. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to non-INSTANTIATION sentences. First sentences of INSTANTIATION are significantly shorter, have fewer rare words and more gradable adjectives. 21

3.2 Part-of-speech tags used significantly ($p < 0.05$) more often in the first sentences of INSTANTIATION compared to the second ($s_1 > s_2$), significantly more often in the second sentences of INSTANTIATION compared to the first ($s_2 > s_1$), and significantly more ($+$) or less ($-$) often compared to non-INSTANTIATION sentences. A dagger ($\dagger$) denotes that for non-INSTANTIATION sentence pairs the difference is significant in the other direction. 21

3.3 Average Jaccard similarity between sentences and their immediate context. Column 1 & 2: the first and second sentences of INSTANTIATION sentence pairs and of non-INSTANTIATION pairs; column 3: the change between columns 1 and 2. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to non-INSTANTIATION sentence pairs. INSTANTIATION sentences are less similar to each other and to their immediate context. 22

3.4 Precision, recall, F-measure and balanced accuracy of identifying INSTANTIATION. *Inst.*: our system; *Vote*: soft voting between our system and Li and Nenkova (2014); *L&N*: Li and Nenkova (2014); *B&M*: Biran and McKeown (2015); *Lin et al.*: Lin et al. (2014); *Brown-concat*: reimplementation of concatenation of Brown clusters as in Braud and Denis (2015). 23

3.5 Supervised learning results for sentence specificity prediction: accuracy, precision, recall and F measure on general sentences, for sentence surface features (SF), word properties (WP), combined shallow features (SF+WP), Brown clusters (BC), word embeddings (WE), and unigrams (Word identity). It is feasible to predict specificity based on cheaply computable features alone; non-sparse representations of lexical information are more suitable for the relatively small training set. 30

3.6 Performance for sentence specificity prediction at the final stage of co-training, for shallow features (*shallow*), Brown clusters and word embeddings (*BC+WE*), the combination of the two (*combined*), unigrams (*word identity*) and Louis and Nenkova (2011a) (*L&N*). An asterisk (*) denotes significant improvement from L&N ($p < 0.05$, sign test). The co-training system performs the best. The performance of word identity features also improves when more data is added. 33

3.7 Examples of general and specific sentences newly labeled during the co-training procedure. 34

3.8 Average numbers of words, percentages of rare words and percentages of gradable adjectives for: each sentence of SPECIFICATION and INSTANTIATION, and of non-SPECIFICATION sentences. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to non-SPECIFICATION sentences. A dagger (†) denotes significant differences compared to the corresponding INSTANTIATION sentence. Characteristics in SPECIFICATION do not stand out as much as in INSTANTIATION. 37

3.9 Part-of-speech tags used significantly ($p < 0.05$) more often in the first sentences of SPECIFICATION compared to the second ($s_1 > s_2$), significantly more often in the second sentences of SPECIFICATION compared to the first ($s_1 < s_2$), and significantly more (+) or less (−) often compared to non-SPECIFICATION sentences. A dagger (†) denotes that for non-SPECIFICATION sentence pairs the difference is significant in the other direction. 38

3.10 Part-of-speech tags used significantly more or less often in INSTANTIATION than in SPECIFICATION. 38

3.11 Average Jaccard similarity between relation sentences and their immediate context before the first sentence, for SPECIFICATION and INSTANTIATION (row 1) and sentences not of the corresponding relation (row 2). The last columns of each relation show the change between s_1 and s_2's similarities. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to sentences not of the relation. SPECIFICATION sentences are more similar to their context than INSTANTIATION. 39

4.1 Number of question interrogatives used by the three annotators and percentages of the context status associated with each question. Largest values in each row are bolded. "How", "why" and "when" questions have stronger association with answers not present in prior context. 50

4.2 Percentages of part of speech tags that are not highlighted (*specified*) and those that are marked as underspecified with associated context status (*immediate, previous, none*). Most of the underspecification are from content words; among them, adjectives, adverbs and verbs have stronger association with answers not present in prior context. 51

4.3 Performances for sentence specificity prediction on the same data as in Table 3.6. *Attn-withnum*: attention network without processing numbers; *attn*: with special symbol for numbers (default); *attn-ne*: with special symbol for named entities; *attn-pron*: with special symbol for unresolved pronouns; *attn-ne-pron*: with both named entity and unresolved pronoun symbols. Precision and recall are on general sentences. First and second best values for each measure are bolded. 58

4.4 Performance for sentence specificity prediction on our annotated data described in Section 4.1. *Attn*: attention network with special symbol for numbers (default); *attn-ne*: with special symbol for named entities; *attn-pron*: with special symbol for unresolved pronouns; *attn-ne-pron*: with both named entity and unresolved pronoun symbols. Precision and recall are on general sentences. 59

4.5 Average and standard deviation of the number of question tokens per sentence, for all question tokens and question tokens that are not-in-context. . 60

4.6 Accuracy and F measure for token specificity prediction, when the number of question tokens is known. Column *all* shows all question tokens; column *N context* shows question tokens that are not-in-context. 60

5.1 Pearson (Spearman) correlation coefficient between lengths of source sentences and HTER values of three MT systems, for Chinese (ZH) and Arabic (AR). There is no strong relationship between sentence length and HTER values. 69

5.2 *Left:* ANOVA with type of segment (1-1 or 1-many) as independent variable and the three MT systems as subjects. *Right:* average HTER values for the three Chinese to English systems for 1-1 and 1-many segments. An asterisk ($*$) denotes significance at $p < 0.05$. 1-many segments is a significant factor in Chinese to English MT quality. 71

5.3 Number of discourse connectives and MT quality. An asterisk (∗) or a plus (+) sign denotes significance at 95% and 90% confidence levels, respectively. Using more than one connective vs. no connective is a significant factor in Chinese to English MT quality. 72
5.4 Relation sense and MT quality. An asterisk (∗) or plus (+) sign denotes significance at 95% and 90% confidence levels, respectively. The presence of a CONTINGENCY relation is a significant factor in Chinese to English MT quality; the interaction between COMPARISON and TEMPORAL is significant for both Chinese and Arabic. 73
5.5 The impact of discourse connective mismatch between human edits and system translations on MT quality, for 1-1 and 1-many segments. An asterisk (*) denotes significance at $p < 0.05$. This mismatch is a significant factor in MT quality from both Chinese and Arabic to English. 75
5.6 Examples of Chinese sentences expressed in multiple English sentences. . . 77
5.7 Percentage of Chinese sentences for which a given number of translators (# ref multi) prefer to use multiple sentences in English (% data), along with percentage of times a multi-sentence translation was selected as most fluent and comprehensible by readers (% best multi). 79
5.8 Percentage of content-heavy Chinese sentences, along with BLEU scores for heavy and non-heavy sentences and their difference. The BLEU score for content-heavy sentences are much lower. 81
5.9 Counts of heavy (Y) and non-heavy (N) sentences with and without full-stop commas. 81
5.10 Performance to identify heavy sentences using multiple reference data (parallel) vs. full-stop comma oracle labels (oracle comma) and predicted full-stop commas (predicted comma). It is more advantageous to learn from multiple reference data. 83
5.11 Accuracy, precision and recall of classifying content-heavy sentences, using MTC and/or OpenMT as training data. *Baseline*: sentence length; *full set*: full set of features proposed in our work. 86

5.12 Number of segments, precision, recall and posterior probability for examples where at least 0, 1, 2, 3 or 4 translators split the sentence. When more translators split the sentence, the classifier is more confident and achieves better performance. 87

5.13 For each criterion to separate heavy and non-heavy sentences, the percentage of heavy sentences (*%data(Y)*), BLEU scores for heavy and non-heavy sentences, and their differences. The criteria are: *fs-comma*: whether the sentence contains a full-stop comma; *length threshold*: whether it is longer than the length threshold; *pred-heavy (prob)*: whether it is predicted predicted content heavy with the posterior probability cutoff *prob*; *oracle heavy*: whether it is content heavy according to the oracle definition. Our system can more reliably identify sentences that are harder to translate. 89

6.1 Average specificity of heavy (H) and non-heavy (¬H) reference translations given a length range of the source sentence: for all references (*all*), for only one-sentence translations (*excl_multi*), when all references are treated as one-sentence translations (*all, as-one-sent*), and the corresponding specificity per-token (*token specificity*). Bold font means significance ($p < 0.05$) when compared to non-heavy sentences. Content-heavy sentences have higher per-token specificity but translating them into multiple sentences leads to lower overall specificity. 95

6.2 Patterns of general and specific sentences within each multi-sentence translation among the heavy Chinese sentences. The least frequent pattern is general-specific, hence an INSTANTIATION relation is not likely to hold between two split components of a heavy sentence. 96

6.3 Number of sentence pairs with an implicit INSTANTIATION per 200 words in Chinese and English. The PDTB-Gold column shows the numbers for gold-standard annotations in the PDTB; others are predicted with the system in Section 3.2. The rate of predicted implicit INSTANTIATION is much lower in Chinese. 97

6.4 Specificity of identified implicit INSTANTIATION arg_1s vs. arg_2s for 5 random draws among the 4 reference translations. The second arguments of identified INSTANTIATION are significantly less specific. 98
6.5 Counts of predicted INSTANTIATION arguments that are content-heavy, along with percentages of such sentences among all heavy sentences, among all arg_1s and among all arg_2s. There is a strong association between content-heavy sentences and arg_2s of INSTANTIATION. 98
6.6 Average counts (#) and average percentages (%) of explicit discourse relations whose arguments are: between whole Chinese sentences, between split components within a sentence and within a split component. Larger percentages in each row are bolded. The distribution of discourse relations across split components is very different from that of relations that do not trigger a split. 99
6.7 Number of predicted implicit INSTANTIATION relations within multi-sentence translations, along with its percentages among heavy sentences, for each translator separately (*Ref1-4*) and averaged among all translators (*avg*). Implicit INSTANTIATION is not a strong trigger for splitting heavy sentences. . 100
6.8 Accuracy, precision, recall and F-measure for content heavy sentence prediction. *P*: posterior probability from our content-heavy classifier in Section 5.2; *L*: sentence length; *S*: specificity; *ST*: per-token specificity; *I*: probability of being a predicted INSTANTIATION arg_2. Specificity information complements L and P in predicting heavy sentences. 101

7.1 Percentages of original-simplified sentence pairs with lower attribute values for the simplified side (%pairs), along with mean values for each attribute among simplified and original sentences. Specificity of simplified sentences are remarkably lower. 105
7.2 Precision for identifying sentences to simplify. Specificity and ARI outperform other attributes. 106
7.3 Spearman correlation for the attributes in original sentences. The correlation between specificity and ARI are not very high. 106

7.4 Numbers of male and female subjects, 95% confidence intervals and mean/standard deviation of specificity perception values. Male subjects rate sentences significantly less specific than female subjects. 111
7.5 Cutoff AQ scores for the high-AQ group and number of subjects for control (G1) and high-AQ (G2) groups. 112
7.6 Spearman correlations of specificity perception values and AQ scores, average perception values among subjects for each group, and fractions of raw sentence ratings outside 95% confidence intervals of crowd ratings. There is a non-significant trend that subjects with higher AQ scores give more specific ratings. 112
7.7 Numbers of male and female subjects, 95% confidence intervals and mean/standard deviation of summary specificity. There is a non-significant trend that female subjects tend to write less detailed summaries. 113
7.8 Spearman correlation of summary specificity and AQ scores, average summary specificity ratings among subjects for each group. There is a non-significant trend for Article 2 that subjects with higher AQ scores tend to write less detailed summaries. 113

List of Illustrations

2.1 An example of ELABORATION and NARRATION from Lascarides and Asher (2007). 10

3.1 Co-training accuracies of sentence specificity prediction with increasing number of unlabeled sentences. 32

3.2 Scatter plot of sentence length (x-axis) and posterior probabilities of sentence specificity prediction (y-axis) on the test data. Our system gives more probabilities towards 0 and 1, hence the higher correlation with sentence length than Louis and Nenkova (2011a). 35

4.1 Distribution of sentence specificity ratings among the three annotators. . . 46

4.2 Average fraction of tokens marked as underspecified vs. average sentence specificity ratings. 48

4.3 Architecture of the attention network for sentence specificity prediction. . . 55

4.4 Precision at 3, 6, 9, 12 tokens for token specificity prediction. X-axis: number of tokens; y-axis: precision. 61

5.1 Commas separating coordinating IPs at the root. These full-stop commas can act as sentence boundaries in Chinese (Xue and Yang, 2011). 82

6.1 Histograms of pairwise specificity differences of reference translations of the same Chinese sentence; showing only pairs where both translations are single-sentence or multi-sentence. X-axis: pairwise specificity difference; y-axis: number of sentence pairs. Red lines indicate the average difference between 1000 randomly selected sentences. For the same source sentence, the specificity of translations of the same type (either one- or multi-sentence) is much more consistent compared to those of different types (Figure 6.2). 93

6.2 Histograms of pairwise specificity differences of reference translations of the same Chinese sentence; showing only pairs where one translation is single-sent and the other is multi-sent. X-axis: pairwise specificity difference; y-axis: number of sentence pairs. Red lines indicate the difference between 1000 randomly selected sentences. For the same source sentence, the specificity of translations of different types (one- or multi-sentence) is much more different compared to those of the same type (Figure 6.1). 94

7.1 Boxplots of AQ scores among subject groups. Red line: median; red dot: mean; green arrows: 95% confidence intervals; boxes/whiskers: 75%-95% quantiles. AQ scores do not differ significantly across gender but differs for non-native speakers. 110

7.2 Distribution of AQ scores. Vertical lines indicate cutoff AQ scores of the control and high-AQ groups (red: female, blue: male). 112

Chapter 1

Introduction

To communicate effectively through language, a writer needs to organize the content to be conveyed into intelligible text that flows naturally. Often, ways of organization that are considered proper among one group of audience are not conventional for another audience. Consider the following example:

> The Dutch, under the leadership of Jan Pieterszoon Coen, captured and razed the city in 1619, after which the capital of the Dutch East Indies — a walled township named Batavia — was established on the site. (*Encyclopedia Britannica*)
>
> The Dutch captured and destroyed the city in 1619. They then constructed a new town and named it Batavia. (*Britannica Elementary*)

In addition to the lexical and syntactic simplifications usually modeled in automatic text simplification systems (Siddharthan, 2014), the authors of Britannica Elementary selectively removed details, made some generalizations, and reorganized the content into two different sentences. Consequently, the text is more accessible to children.

Furthermore, the use of discourse devices that organize information into complex sentences differ across languages:

> 来自美国、日本、新加坡的外资增加较多，新项目中外商投资比例越来越高，独资企业明显增加。**[literal]** *From U.S., Japan, Singapore foreign investment increase more, new projects among foreign funds proportion higher and higher, solely foreign enterprises considerably increase.*

> [**Translation**] The foreign investment from U.S., Japan, and Singapore has increased more. In new projects, the proportion of foreign funded ones is getting higher and higher, and wholly foreign owned enterprises have considerably increased.

The single Chinese sentence is expressed in two English sentences. The CONJUNCTION relation, signaled by the discourse connective "and", is expressed implicitly in Chinese. Again, these differences are separate from lexical choice and syntactic transformations that Machine Translation systems are more adept at.

The above phenomena demonstrate an interplay between *the amount of detail in text*, *the organization of information into sentences*, as well as *the expression of discourse relations*. This thesis presents techniques and insights that capture factors among the three discourse aspects that play significant roles in proper text understanding and communication. We study how audiences react to text in this respect among broadly applicable groups: readers who speak different languages, have lower reading ability (e.g., children, language learners), or have impaired communication ability.

Well-organized text involves careful arrangements of general statements and details, as well as decisions about the appropriate amount of detail to express, i.e., the level of *specificity*. For example, in the news snippet below, the first sentence invokes a reader's interest by drawing a general picture of the situation while the second sentence supplies details:

> Evidence of widespread cheating has surfaced in several states in the last year or so. California's education department suspects adult responsibility for erasures at 40 schools that changed wrong answers to right ones on a statewide test.

The organization of general and specific content follows patterns that ensure both the main purpose of the text and the interpretation of details are efficiently and correctly communicated (Dixon, 1987), hence impacting the coherence and the quality of text (Scanlan, 2000; Higgins et al., 2004; Louis and Nenkova, 2013b). In addition, different target audiences (e.g., non-experts vs. experts) can vary in their perception of specificity and the amount of detail they are comfortable to comprehend. Text that lacks specificity also often relies on context to be fully comprehended. This thesis contributes to the understanding of text

specificity and efficiency and scalability in specificity prediction, making this property practically accessible for other applications and studies. We also identify lexical and coherence characteristics concerning the flow of general and specific content in adjacent sentences, in particular, those that have an INSTANTIATION discourse relation.

We study differences in the packaging of general and specific content into sentences across two languages, Chinese and English. We found that a significant portion of Chinese sentences needs to be reorganized into multiple English sentences for proper intelligibility (we name them *content-heavy sentences*). Sentence length is not a sufficient indicator to correctly decide if a sentence is content-heavy. These sentences are also associated with higher specificity per-word. We develop a novel system to accurately identify content-heavy sentences. We further present compelling evidence that without properly handling discourse-related variabilities across languages, the quality of translated text can be affected. It can be especially problematic for Machine Translation (MT) systems, which normally translate a sentence in one language into a single sentence in another.

Within the same language, our studies indicate that the perception of specificity is influenced by an individual's gender and traits. We found that a decrease in the level of specificity is a steady characteristic in human simplification that targets language learners and children. We also discovered links between gender and a reader's perception of text specificity. Finally, we performed a pilot study, investigating text specificity perception and production in adults with varying scores in a diagnostic test for the Autism Spectrum Disorder (ASD), a neuro-developmental disorder characterized by impaired ability in verbal and non-verbal communication. The preliminary results indicate weak trends and reveal necessary adjustments in the experimental design for future studies.

1.1 Thesis contributions

In this thesis, we present novel techniques for text specificity prediction and cast new insight into the packaging of general and specific information in discourse structure. We show that different groups of target audiences follow distinct conventions in discourse organization and diverge in text specificity perception, and that it is possible to identify these conventions to improve the intelligibility of text.

Techniques for text specificity prediction. We present SPECITELLER (Section 3.3), a semi-supervised system for sentence specificity prediction. SPECITELLER targets three key aspects of improvement from traditional approaches. It uses lightweight text processing, scalable representations for words and bypasses expensive human annotation. It relies only on the surface text string with no overhead from costly analysis such as syntactic parsing and named entity recognition. SPECITELLER yields significant improvement over prior work and is currently the only system publicly released to predict sentence specificity.

We also present the first work to predict tokens within a sentence that need further elaboration (Section 4.2). Our system is trained for sentence specificity prediction and ranks token specificity in an unsupervised manner using an attention network. We show promising results and discuss practical steps for extensions.

Insights into specificity and discourse structure. The lack of specificity can be recognized through various discourse factors, for example, referring expressions to entities or events, adjectives with previously established degrees, expressions to get a reader's attention, etc. We develop an annotation schema that enables systematic analysis on the degree of sentence specificity, the location of underspecified expressions and the reason for the lack of specificity (Section 4.1). We find that lack of specificity impacts text understanding by frequently invoking high-level comprehension questions among readers.

We further reveal distinct lexical and coherence characteristics associated with the INSTANTIATION discourse relation, which set it apart from all other relations and make it an ideal training source for sentence specificity. Additionally, features capturing these characteristics alone result in an improvement of 8% in accuracy when predicting the relation (Section 3.2).

Coherence preferences in cross-lingual communication. We identify discourse factors that significantly impact the quality of translations from Chinese to English, challenging the traditional sentence-to-sentence view of machine translation systems. A large portion of Chinese sentences can lead to poor understanding when translated as a single sentence into English, which we define as *content-heavy*. We observe that these sentences are more specific comparing to non-heavy sentences of similar lengths (Section 6.2). We present a

novel system to effectively identify these sentences and show that the predicted sentences are associated with substantially worse MT performance (Section 5.2). We further identify other discourse factors, such as the use of discourse connectives, that are different in the two languages and can significantly impact MT quality (Section 5.1). This is the first work in this direction for translation from Chinese to English.

Coherence preferences in monolingual communication. We investigate the influence of three group characteristics in text specificity perception: gender, reading abilities and autism traits. We found that sentence specificity is an effective indicator of sentences that are hard to understand and need simplification. The specificity of simplified sentences is lower, suggesting that beginner readers are more comfortable with fewer details (Section 7.1).

We design and conduct the first study that investigates the connection between the perception and production of text specificity among male and female, and people with varying autism-like symptoms. We observe significant gender differences in specificity perception. Further, there are non-statistically significant tendencies among people with more autism-like symptoms to rate sentences to be more detailed than others would, and to produce less detailed summaries for popular science articles (Section 7.2).

Publicly available resources. We release the following tools and datasets:

- SPECITELLER, a fast and accurate semi-supervised sentence specificity predictor: `http://www.cis.upenn.edu/~nlp/software/speciteller.html`
- Corpus of sentence specificity ratings and underspecified segments along with their locations in text: `http://www.cis.upenn.edu/~nlp/corpora/lrec16spec.html`

1.2 Thesis organization

Chapter 2 presents an overview on concepts and systems related to text specificity. The chapter covers three main topics: (a) discourse relations, (b) the organization of general and specific information, and (c) notions of specificity in fine-grained units.

Chapter 3 discusses the connection between discourse relation and sentence specificity prediction. An implicit INSTANTIATION relation between two sentences signals that the second sentence discusses in further details some aspects of the content in the first sentence. Previously this relation was shown to have significantly less entity overlap between its two arguments (Louis and Nenkova, 2010). We discover that this lack of coherence extend to smaller overall word overlap between the two sentences. Instead, patterns in word usage are highly distinctive and set apart the relation's first sentence and the sentence pair from all other discourse relations, including the one that is closest in definition, SPECIFICATION. Exploiting these characteristics substantially improves the prediction of the relation. Furthermore, INSTANTIATION's characteristics can be extended and generalized to predict specificity in all sentences by means of bootstrapping, resulting in our highly effective system SPECITELLER.

Chapter 4 presents our annotation and corpus study on the connection between the lack of specificity within a sentence and prior context. The annotation scheme marks underspecified expressions within a sentence and records how they can be specified via a question-answering exercise. We found that a third of the expressions marked underspecified are not elaborated anywhere in prior context, and that they are much more likely to trigger high-level questions such as "how" and "why". This chapter also describes the first system to predict these marked tokens within a sentence. The system learns to predict sentence specificity and ranks words using an attention network.

Chapter 5 describes divergences in discourse organization between Chinese and English that highly impact the quality of Machine Translation from Chinese to English. One of the most notable divergences is the mismatch of the acceptable amount of information in one sentence between the two languages. We define *content-heavy* sentences to be Chinese sentences whose content is too much to be packed into one English sentence and describe our system to reliably identify such sentences. We also identify differences in the expression of discourse relations, such as the use of connectives, that also influence MT quality significantly.

Chapter 6 discusses the three aspects of organizing information studied in the previous chapters: the discourse relation INSTANTIATION, text specificity, and content-heavy sentences. We show that content-heavy sentences in Chinese are associated with higher specificity per-word and share similar properties with the INSTANTIATION relation in English. The two aspects are complementary to the syntactic characteristics of heavy sentences explored in Chapter 5. Their translations into multiple English sentences are further associated with a distinct distribution of explicit discourse relations involved in sentence splitting, demonstrating additional differences in content organization between the two languages.

Chapter 7 presents two studies that investigate specificity and characteristics of readers. First, we target beginner readers such as children and language learners by studying data for sentence simplification. We find that specificity for simplified sentences is lower, and that sentence specificity is an effective indicator in determining whether a sentence needs to be simplified. Second, this chapter reports findings from our study which examines associations between each subject's gender and assessments for autism-like symptoms, and their sentence specificity judgements and specificity of their summaries of news articles.

Chapter 8 concludes the thesis and lays out future directions.

Chapter 2

Notions of specificity in text

Text specificity discussed in this thesis captures the level of details in text. This broad definition is motivated by the necessity to properly organize general and specific content in discourse. We develop tools to harness text specificity so that they can be used for tasks such as recognition of important information, characterizing readability, and assessing text quality.

Our work is related to several existing notions of specificity in semantics and pragmatics, for example, certain discourse relations, generic expressions, and entity instantiation. They capture specificity of text from different and more specific angles. In later chapters, we will look back to these concepts and elicit in more detail how our work is related.

First, we discuss discourse relations that indicate changes in the level of details in text: ELABORATION and INSTANTIATION. This is most related to our work, since we used annotated INSTANTIATION as training data for sentence specificity prediction (Chapter 3). Although these relations are well-defined and annotated, we point out that current work is limited in linking text specificity and discourse relations, and motivate our work to characterize INSTANTIATION.

We review prior studies that show how proper flow of specificity in text and in conversations is essential for effective communication. In our work, we examine the cause and effect of the lack of specificity when reading an article (Chapter 4), and complement prior studies by developing annotation guidelines and tools to automate word specificity prediction given a sentence.

Constraints on the N+S combination: S presents additional detail about the situation or some element of subject matter which is presented in N or inferentially accessible in N in one or more of the ways listed below. In the list, if N presents the first member of any pair, then S includes the second:
1. set : member
2. abstract : instance
3. whole : part
4. process : step
5. object : attribute
6. generalization : specific

The effect: Reader recognizes the situation presented in S as providing additional detail for N. Reader identifies the element of subject matter for which detail is provided.
Locus of the effect: N and S

Table 2.1: Definition of ELABORATION in Rhetorical Structure Theory (Mann and Thompson, 1988).

INSTANTIATION: arg_1 evokes a set of events and arg_2 picks up one of these events and describes it in further detail. In logical terms, we have $exemplify'(arg_2, \lambda x.x \in g(arg_1))$ where g is a function that extracts the set of events from the semantics of arg_1, x is a variable ranging over them, $examplify'$ asserts that arg_2 further describes one element in the extracted set.

SPECIFICATION: The semantics of arg_2 restates the semantics of arg_1 and the situations described in arg_1 and arg_2 hold true at the same time. Further, $arg_1 \leftarrow arg_2$ where $\leftarrow$ denotes logical implication.

Table 2.2: Definition of INSTANTIATION and SPECIFICATION in the Penn Discourse Treebank (Miltsakaki et al., 2008).

Finally, we discuss other, more specific notions of text specificity which applies to clauses and phrases: generic expressions and entity instantiations. We point out how prior work in these areas are related to ours. Additionally, we distinguish the senses of "specificity" and "underspecification" used in this thesis from those in semantics.

2.1 Discourse relations

The flow of general (less detailed) and specific (more detailed) content in text is often captured by certain discourse relations. The ELABORATION relation in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), defined as in Table 2.1, clearly demonstrates a change in specificity across the two text spans it connects. In the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), the INSTANTIATION and SPECIFICATION relations—defined as in Table 2.2—also account for specificity changes in a similar manner. Among these two

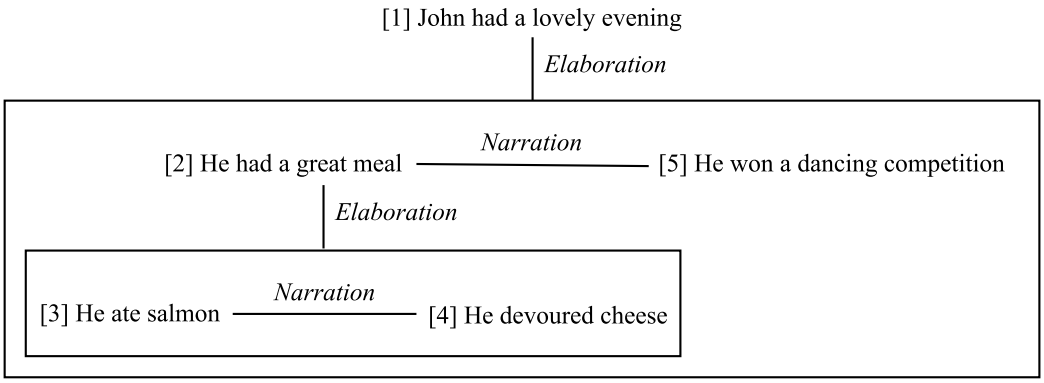


Figure 2.1: An example of ELABORATION and NARRATION from Lascarides and Asher (2007).

relations, INSTANTIATION shows clearer characteristics of general and specific information and is the better choice to be used as training data for sentence specificity prediction. We will discuss in detail the difference between the two relations in Section 3.4.

Consider the example in Figure 2.1 from Lascarides and Asher (2007). Each ELABORATION gives rise to more detailed descriptions[1]; e.g., the great meal and winning the competition to further explain "a lovely evening", "salmon" and "cheese" to further explain "a great meal". In contrast, the NARRATION relation, which "reflects temporal progression between events" (Lascarides and Asher, 2007), does not capture specificity changes across their connected text spans.

Attempts to detect discourse relations fall under the rubric "discourse parsing", but identifying these relations, particularly for the purpose of harnessing specificity, is little explored. Current RST discourse parsers treat ELABORATION as a single class during training and prediction (Ji and Eisenstein, 2014; Feng and Hirst, 2014; Joty et al., 2015); however, the most common subclass ELABORATION-ADDITION, which was added in the RST Discourse Treebank, does not necessarily signify a change in specificity (Carlson and Marcu, 2001), e.g.,

> *Under a proposal by Democrats to expand Individual Retirement Accounts, a $2,000 contribution by a taxpayer in the 33% bracket would save $330 on his taxes.* **The savings was given incorrectly in Friday's edition.**

[1]Note that the definition of ELABORATION in Lascarides and Asher (2007) is: "ELABORATION(π_1, π_2) entails that the events described in π_2 describe in more detail those described in π_1". This definition is not the same as in RST.

The same issue exists for INSTANTIATION: many state-of-the-art PDTB parsers only identify EXPANSION, which subsumes INSTANTIATION with other relations irrelevant for specificity (e.g., CONJUNCTION) (Rutherford and Xue, 2014; Braud and Denis, 2015; Liu and Li, 2016; Ji et al., 2016). More importantly, about 83% of the time, an INSTANTIATION relation is not signaled by cues such as "for example" and needs to be inferred from adjacent sentences. Identifying implicit discourse relations is a well-known difficult task, with best F-measures around 0.4 across the standard set of relations in CoNLL shared tasks (Xue et al., 2015; Xue et al., 2016).

One approach that can be applied particularly to ELABORATION and INSTANTIATION utilizes textual entailment. In Rus et al. (2009), the ELABORATION relation is treated as a type of textual entailment where arg_2 entails arg_1 by providing more material to arg_1, hence the relation can be recognized with a graph-based entailer if arg_1 is subsumed in the graph of arg_2. However, we found that out of the 1,457 instances of implicit INSTANTIATION in the PDTB, only 20 were tagged entailment with the Excitement Open Platform for Recognizing Textual Entailments (Magnini et al., 2014). Hence identifying INSTANTIATION with this approach is not reliable. In Section 3.2.4, we will explore in detail why the recognized entailment rate in INSTANTIATION is low.

Although there is clear overlap between specificity and discourse relations, prior work at the intersection of the two is limited. In this thesis, we focus on characterizing INSTANTIATION and utilizing the relation for sentence specificity prediction (Chapter 3).

2.2 Information organization

Prior work has revealed that the organization of general and specific content is important for achieving communication goals. Dixon (Dixon, 1982; Dixon, 1987) considered the order of presentation of two types of information in procedural directions: more general, organizational information (such as [a] below) and the specific step descriptions (such as [b] below):

[a] This will be a picture of a wagon.

[b] Draw a long rectangle with two circles underneath.

He showed that reading time is shorter when organizational information was presented before the step descriptions. He further showed that by presenting the more specific, step descriptions first will cause the reader to mentally make guesses about the high-level picture, making reading more effortful. This indicates that the proper flow of general information and details helps reading comprehension.

At the same time, in natural, effective communication, the specificity of utterances vary as more context is established. Djalali et al. (2011) studied conversations between two players in a domain-restricted card game under the Question Under Discussion (QUD) framework (Beaver and Clark, 2009; Roberts, 1996). A QUD is the current implicit or explicit question that the interlocutors are set to resolve, given the common ground they share. Questions vary in specificity, as shown below from general to specific:

Depth 1 How do I interact with the game world?

Depth 3 What is the expertise of my fellow player?

Depth 7 Have we obtained a particular winning sequence?

Djalali et al. (2011) gathered 439 chat histories of gameplay and categorized each utterance into 7 depths. They showed that the more expertise the players had with the game, the more specific questions they asked at the start of the game. In other words, the more gaming context the players had established, the more specific their interchanges were.

Our work seeks to understand how general, underspecified information impacts comprehension as one reads an article; i.e., what questions readers may have in mind when they encounter such information. Contrary to prior approaches, we carry out our analysis with the goal of developing annotation guidelines, datasets and systems for automatic specificity prediction of text within a sentence. In our pilot annotation for text specificity (Section 4.1), we explore the effect of the lack of specificity on reading comprehension as a question-answering exercise. We show that the lack of specificity triggers high-level questions. We also present the first system to predict word specificity given a sentence (Section 4.2). Furthermore, we find that the cause of lack of specificity changes as readers get more into an article, aligning with findings in Djalali et al. (2011). We will discuss this in more detail in Section 4.1.4 ("context status and sentence number").

2.3 Fine-grained linguistic structure

Generics and habituals. Prior work has explored two important phenomena related to general and specific information: whether an expression describes a class of entities instead of specific individuals (i.e., *generics*), and whether an expression describes regularities rather than specific events (i.e., *habituals*) (Carlson, 2005):

- **generic:** Sugar maples also have a tendency to color unevenly in fall. (Friedrich and Pinkal, 2015)
- **non-generic:** Potatoes are on the kitchen counter.
- **habitual:** After 1971 Paul Erdos also took amphetamines. (Reiter and Frank, 2010)
- **non-habitual:** Paul Erdos was born [...] on March 16, 1913. (Reiter and Frank, 2010)

Computational approaches to identify generics and habituals adopt a variety of statistical models with elaborate feature engineering, including Decision Trees and Naive Bayes (Mathew and Katz, 2009), Maximum Entropy (Palmer et al., 2007), Bayes Net (Reiter and Frank, 2010) and Conditional Random Fields (Friedrich and Pinkal, 2015). Reiter and Frank (2010) found that certain word usages such as numbers and plural nouns are among the most indicative features for classifying generics. This echoes our finding that numbers are more likely to associate with specific sentences and numbers with general sentences (Section 3.2).

With our tools developed for predicting text specificity, we found that habituals, generic clauses, and clauses that are led by generic noun phrases are significantly less specific per-word than others. We discuss this in detail in Section 3.3.6.

Entity instantiation. Entity instantiation is a type of entity relation "in which a set of entities is introduced, and either a member or subset of this set is mentioned afterwards" (McKinlay and Markert, 2011). By jumping from a set to a member or a subset, entity instantiations also signal a change in specificity. Some examples from McKinlay and Markert (2011) are shown below:

[set-member] **Some European funds** recently have skyrocketed. *Spain Fund* has

surged to a startling 120% premium.

[set-subset] **Bids totalling $515 million** were submitted. *Accepted offers* ranged from 8.38% to 8.395%.

McKinlay (2013) studied the co-occurrence between discourse relations and entity instantiation across the two arguments of a relation, using their own annotation over part of the PDTB corpus. They found that the INSTANTIATION relation has significantly more overlap with entity instantiations than other discourse relations. In our work, we explore the detection of the INSTANTIATION relation, which signals a change between two sentences instead of phrases.

Referents and underspecification in semantics Finally, we would like to differentiate our use of the terms "specificity" and "underspecification" from that in semantics. In semantics, specificity refers to "the degree of individuation of an entity" or "the uniqueness of an entity" (Frawley, 1992). A specific noun phrase refers to a particular entity that the speaker has in mind in a given context; a non-specific noun phrase refers to a class instead of an individual entity. In our work, we use "specificity" to refer to the level of details in text.

"Underspecification" in semantics refers to a technique used when the interpretation of the meaning of a proposition is ambiguous. Underspecification uses one single representation to include all possible readings given by the ambiguity (van Deemter and Peters, 1996). This is quite different from our work; we refer to a text segment as underspecified if the reader needs elaboration or explanation of the segment in order to fully understand its sentence.

Chapter 3

From discourse relation to text specificity

Discourse relations are semantic relations between text spans. For example, sentences S_1 and S_2 below have an INSTANTIATION relation, with S_2 elaborating on one of the cases in which McDonald's needed more eggs:

> S_1: With the national announcement last week of plans to sell some breakfast items all day long, the company expects to buy even more eggs.
>
> S_2: For example, the Egg McMuffin, which uses one egg per sandwich, is among the company's most popular menu items.

Researchers have demonstrated the value of discourse relations in a number of natural language processing tasks, for example, content selection in abstract and compressive summarization (Hirao et al., 2013; Kikuchi et al., 2014; Gerani et al., 2014; Durrett et al., 2016), assessing machine translation quality (Li et al., 2014; Guzmán et al., 2014), question generation and answering (Chai and Jin, 2004; Prasad and Joshi, 2008; Agarwal et al., 2011) and sentiment analysis (Bhatia et al., 2015; Hogenboom et al., 2015).

Text specificity on the other hand is a property associated with a single textual unit

Content in Section 3.2 is published at NAACL 2016 (Li and Nenkova, 2016). Content in Section 3.3 published at AAAI 2015 (Li and Nenkova, 2015b). We thank Ariani Di Felippo for her contribution to the analysis of Section 3.2.4.

(words, phrases, sentences, etc.). Clearly written texts tailor the specificity of content to the intended reader and exhibit clear patterns in the flow of specificity (Scanlan, 2000; Higgins et al., 2004). Consider the following two sentences, talking about test cheating:

> [**general**] Evidence of widespread cheating has surfaced in several states in the last year or so.
>
> [**specific**] California's education department suspects adult responsibility for erasures at 40 schools that changed wrong answers to right ones on a statewide test.

Both sentences convey the information that (exam) cheating is taking place. The first sentence is rather general: it contains a vague piece of information on the extent of cheating (*widespread*), the location of cheating (*several states*) and time of the cheating (*in the last year or so*) and says nothing about exactly what cheating consisted of. The second is more specific, conveying that it was not students who did the cheating and that 40 schools in California were suspected and what activities constituted the cheating, making the extent, location and exact events much more precise.

The specificity of text is expressed on multiple levels, and can be quantified at the level of words (*people* vs. *students* vs. *Mary Smith*) (Reiter and Frank, 2010; Krahmer and van Deemter, 2012), sentences, as in the example above (Mathew and Katz, 2009; McKinlay and Markert, 2011), or full texts or paragraphs, where the distinction boils down to determining if the intended audience of a text are lay people or experts (Elhadad et al., 2005). In practice, Louis and Nenkova (2011a) showed that changes in sentence and overall text specificity are strongly associated with perceptions of text quality. Science writing of the best quality in the New York Times is overall more general than regular science pieces in NYT and contain fewer stretches of specific content (Louis and Nenkova, 2013a). Automatic summaries, which are often judged to be incoherent, are significantly more specific than same length human-written summaries for the same events (Louis and Nenkova, 2011b). It is also a stable predictor in identifying high-quality arguments in online discussions (Swanson et al., 2015) and in characterizing informativeness in political discourse (Cook, 2016).

In this chapter we focus on the the discourse relation INSTANTIATION and its important role in deriving sentence level specificity. We catalog characteristics that set apart INSTANTIATION from all other discourse relations, including the one that is the closest in

definition, SPECIFICATION. Such information not only gives rise to the substantial improvement of predicting this mostly-implicit relation, but also can be successfully generalized in unlabeled dataset to derive sentence-level specificity with high accuracy. In our sentence specificity prediction system SPECITELLER, we forgo costly processing such as syntactic parsing and named entity recognition, and extract lightweight string surface features with dictionary lookups. SPECITELLER achieves significant improvement over prior work (Louis and Nenkova, 2011a) that uses a much more complicated set of features.

3.1 Discourse relations in the Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) contains annotations for five types of discourse relations over the Penn Treebank corpus (Marcus et al., 1993). It is by far the largest lexical based discourse relation annotation corpus.

In the PDTB, discourse relations are viewed as a predicate with two arguments. The predicate is the relation, the arguments correspond to the minimum spans of text whose interpretations are the abstract objects between which the relation holds. Consider the following example of a contrast relation. The italic and bold fonts mark the arguments of the relation.

> Commonwealth Edison said *the ruling could force it to slash its 1989 earnings by $1.55 a share*. [Implicit = BY COMPARISON] **For 1988, Commonwealth Edison reported earnings of $737.5 million, or $3.01 a share**.

For explicit relations, the predicate is marked by a discourse connective that occurs in the text, e.g., *because*, *however*, *for example*.

Implicit relations are annotated between adjacent sentences in the same paragraph. They are inferred by the reader but are not lexically marked. Alternative lexicalizations (*AltLex*) are the ones where there is a phrase in the sentence implying the relation but the phrase itself is not one of the explicit discourse connectives. The annotators were asked to come up with a connective that *could have been inserted* to connect the two sentences. There are 16,224 and 624 examples of implicit and *AltLex* relations, respectively.

Relation senses in the PDTB form a 3-level hierarchy. Annotators were asked to iden-

tify the most fine-grained relation, and then backed off to one level higher when there was disagreement. The top level relations are COMPARISON (arg_1 and arg_2 holds a contrast relation), CONTINGENCY (arg_1 and arg_2 are causally related), EXPANSION (arg_2 further describes arg_1) and TEMPORAL (arg_1 and arg_2 are temporally related). Some of the largest second-tier relations are under EXPANSION, including CONJUNCTION (arg_2 provides new information to arg_1), INSTANTIATION (arg_2 exemplifies arg_1) and RESTATEMENT (arg_2 semantically repeats arg_1).

Finally, 5,210 pairs of adjacent sentences were marked as related by an entity relation (*EntRel*), by virtue of the repetition of the same entity or topic. *EntRel*s were marked only if no other relation could be identified and they are not considered a discourse relation, rather an alternative discourse phenomenon related to entity coherence (Grosz et al., 1995). There are 254 pairs of sentences with no discourse relation identified.

Pitler et al. (2008) showed that accuracy as high as 93% can be easily achieved predicting explicit relations, largely because the connective itself is a highly informative feature. Efforts in identifying the argument spans have also yielded high accuracies (Lin et al., 2014). However, in the absence of a connective, recognizing non-explicit relations has shown to be a real challenge (Liu and Li, 2016; Xue et al., 2016; Braud and Denis, 2015; Ji and Eisenstein, 2015; Xue et al., 2015; Rutherford and Xue, 2014; Biran and McKeown, 2013; Park and Cardie, 2012; Lin et al., 2009; Pitler et al., 2009). In our work of text specificity, the INSTANTIATION relation is the most related. It is predominantly implicit.

3.2 Recognizing INSTANTIATION

In an INSTANTIATION relation, one text span explains in further detail the events, reasons, behaviors and attitudes mentioned in the other (Miltsakaki et al., 2008), as illustrated by the segments below:

> **[a]** Other fundamental "reforms" of the 1986 act have been threatened as well.
>
> **[b]** The House seriously considered raising the top tax rate paid by individuals with the highest incomes.

Sentence **[a]** mentions "other reforms" and a threat to them, but leaves unspecified what

are the reforms or how they are threatened. Sentence **[b]** provides sufficient detail for the reader to infer more concretely what has happened.

The INSTANTIATION relation has some special properties. A study of discourse relations as indicators for content selection in single document summarization revealed that the first sentences from INSTANTIATION pairs are included in human summaries significantly more often than other sentences (Louis et al., 2010) and that being a first sentence in an INSTANTIATION relation is *the* most powerful indicator for content selection related to discourse relation sense. INSTANTIATION is also one of the relations where their first sentences contain more sentiment expressions than other sentences (Trnavac and Taboada, 2013), making the relation useful for sentiment analysis applications. Moreover, INSTANTIATION relations appear to play a special role in local coherence (Louis and Nenkova, 2010), as the flow between INSTANTIATION sentences is not explained by the major coherence theories (Kehler, 2004; Grosz et al., 1995). Many of the sentences in INSTANTIATION relation contain entity instantiations (complex examples of set-instance anaphora), such as "several EU countries"—"the UK", "footballers"—"Wayne Rooney" and "most cosmetic purchase"—"lipstick" (McKinlay and Markert, 2011), raising further questions about the relationship between INSTANTIATIONS and key discourse phenomena.

Detecting an INSTANTIATION, however, is hard. In the PDTB, INSTANTIATION is one of the few relations that are more often *implicit*, i.e., expressed without a discourse marker such as "for example". Of the 1,747 annotated instances, 1,445 of them are implicit. Identifying implicit discourse relation is an acknowledged difficult task, but the challenge is exacerbated due to the lack of *explicit* INSTANTIATIONS: explicit relations are shown to improve their implicit counterparts using data source expansion (Rutherford and Xue, 2015).

We identify a rich set of factors that set apart each sentence in an implicit INSTANTIATION and the pair as a whole. These factors are not word unigrams or their related non-sparse representations but are *patterns* in the usage of words with higher levels of abstraction. We show that these factors improve the identification of implicit INSTANTIATION by at least 5% in F-measure and 8% in balanced accuracy compared to prior systems.

3.2.1 Data and methodology

We use the PDTB for the analysis and experiments presented here. There are a total of 1,747 INSTANTIATION relations in the PDTB, of which 83% are implicit. INSTANTIATION makes up 8.7% of all implicit relations and is the 5th largest among the 16 second-level relations in the PDTB.

We identify significant factors[3] that characterize: *(i)* s_1 and s_2: the first and second sentence in an INSTANTIATION pair vs. all other sentences; *(ii)* s_1 vs. s_2: adjacent sentence pairs in INSTANTIATION relation vs. all other adjacent sentence pairs.

Our analysis is conducted on the PDTB except section 23, which is reserved for testing as in prior work (Lin et al., 2014; Biran and McKeown, 2015). In total, there are 1,337 INSTANTIATION sentence pairs and 43,934 non-INSTANTIATION sentences for the corpus analysis.

3.2.2 Characteristics of INSTANTIATION

Sentence length. Intuitively, longer sentences are more likely to involve details. Table 3.1:#words demonstrates that there is an average of 8.4-word difference in length between the two sentences in an INSTANTIATION relation; moreover, s_1s are significantly shorter (more than 5 words on average) than other sentences, and s_2s are significantly longer.

Rare words. For each sentence, we compute the percentage of words that are not present in the 400K vocabulary of the Glove vector representations (Pennington et al., 2014). Table 3.1:%oov shows that s_1 of INSTANTIATIONs contain significantly *fewer* out-of-vocabulary words compared to either s_2 and non-INSTANTIATIONs. We also compare the difference in unigram probability[4] of content word pairs. Compared to non-INSTANTIATION, words across INSTANTIATION arguments show significantly larger average unigram log probability difference (1.24 vs. 1.22). These numbers show that the first sentences of INSTANTIATION

[3] $p < 0.05$ according to paired Wilcoxon signed rank test for real valued comparison between the two sentences in a relation, non-paired Wilcoxon rank sum for real valued factors in different types of sentences, and Kruskal-Wallis for binary valued features across different types of sentences.

[4] We use a unigram language model on year 2006 of the New York Times Annotated Corpus (Sandhaus, 2008).

#words			%oov			%gradable adj.		
s_1	s_2	¬ Inst.	s_1	s_2	¬ Inst.	s_1	s_2	¬ Inst.
18.4*	26.8*	23.9	0.68*	1.54	1.46	2.96*	2.22	2.22

Table 3.1: Average numbers of words, percentages of rare words and percentages of gradable adjectives per sentence for: first sentences of INSTANTIATION, second sentences of INSTANTIATION, and non-INSTANTIATION sentences. An asterisk (∗) denotes significant differences ($p < 0.05$) compared to non-INSTANTIATION sentences. First sentences of INSTANTIATION are significantly shorter, have fewer rare words and more gradable adjectives.

$s_1 > s_2$	CC EX JJR JJS NNS PDT RB† RBR VBG VBN VBP VBZ†
$s_1 < s_2$	NN NNP† PRP TO VBD WRB
s_1 vs ¬Inst.	CD$^-$ JJ$^+$ MD$^-$ NN$^-$ NNP$^-$ NNS$^+$ PRP$^-$ RB$^+$ TO$^-$ VB$^-$ VBD$^-$ VBG$^+$ VBP$^+$ VBZ$^+$ WDT$^-$
s_2 vs ¬Inst.	CD$^+$ DT$^+$ MD$^-$ NNP$^+$ NNS$^+$ PRP$^+$ RB$^-$ VB$^-$ VBN$^-$

Table 3.2: Part-of-speech tags used significantly ($p < 0.05$) more often in the first sentences of INSTANTIATION compared to the second ($s_1 > s_2$), significantly more often in the second sentences of INSTANTIATION compared to the first ($s_2 > s_1$), and significantly more (+) or less (−) often compared to non-INSTANTIATION sentences. A dagger (†) denotes that for non-INSTANTIATION sentence pairs the difference is significant in the other direction.

do not involve many unfamiliar words — an indication of higher readability (Pitler and Nenkova, 2008).

Gradable adjectives. The use of gradable adjectives (Frazier et al., 2008; de Marneffe et al., 2010)—*popular, high, likely*—may require further explanation to justify the appropriateness of their use. Here we compute the average percentage of gradable adjectives in a sentence. The list of adjectives is from Hatzivassiloglou and Wiebe (2000) and the respective percentages are shown in Table 3.1. Compared to other sentences, s_1 of INSTANTIATION involves significantly more gradable adjectives.

Parts of speech. We study word categories that are heavily or rarely used with INSTANTIATION by inspecting the percentage of part-of-speech tags found in each sentence. In Table 3.2, we show POS tags whose presence is significantly different across arguments in INSTANTIATION but not so across non-INSTANTIATION, with significance in non-INSTANTIATION in the reverse direction denoted by †. Four cases of POS occurrences are inspected:

- more often in s_1 compared to s_2,

	s_1	s_2	Δsim
Inst.	0.0282*	0.0275*	0.0007
$\neg$ Inst.	0.0390	0.0358	0.0042*

Table 3.3: Average Jaccard similarity between sentences and their immediate context. Column 1 & 2: the first and second sentences of INSTANTIATION sentence pairs and of non-INSTANTIATION pairs; column 3: the change between columns 1 and 2. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to non-INSTANTIATION sentence pairs. INSTANTIATION sentences are less similar to each other and to their immediate context.

- more often in s_2 compared to s_1,
- more (+) or less (-) in s_1 compared to non-INSTANTIATION,
- more (+) or less (-) in s_2 compared to non-INSTANTIATION.

We see that s_1 of INSTANTIATION contains more characteristic POS usage than s_2. There are more comparative adjectives and adverbs as well as fewer nouns in s_1 compared to s_2 in INSTANTIATION pairs. The usage of verbs is also different between the two arguments. Compared to other sentences, there are more adverbs and adjectives in s_1, which may be related to findings that they contain more sentiment words (Trnavac and Taboada, 2013). On the other hand, s_2 contains more nouns, numbers, determiners and proper nouns, intuitively associated with the presence of detailed information.

Wordnet relations. We consider word-level relationships across arguments using Wordnet (Fellbaum, 1998). For each noun, verb, adjective and adverb content word pairs across arguments, we calculate the percentage of sentences with each type of Wordnet relation. Among INSTANTIATION sentence pairs there are significantly more noun-noun pairs with hypernym (21.6% vs. 18%) or meronym (18.7% vs. 15.5%) relationships and verbs with indirect hypernym relationship (41.7% vs. 38.5%). We also observe significantly more semantically similar verbs (72.5% vs. 68.7%).

Lexical similarity. Louis and Nenkova (2010) showed that the number of shared entities across sentences having an INSTANTIATION relation is much smaller than other sentences. Here we check lexical similarity in general. We inspect the similarity between sentences in each pair as well as between each sentence in a pair and their immediate prior context; specifically:

System	P	R	F	BA
Inst. specific	0.3072	**0.6986**	**0.4268**	**0.7862**
Vote (L&N)	0.3052	0.6438	0.4141	0.7632
L&N	0.3028	0.4521	0.3626	0.6843
B&M	0.2542	0.2055	0.2273	0.5786
Lin et al.	**0.5500**	0.1507	0.2366	0.5704
Brown-concat	0.1333	0.3836	0.1979	0.5919

Table 3.4: Precision, recall, F-measure and balanced accuracy of identifying INSTANTIATION. *Inst.*: our system; *Vote*: soft voting between our system and Li and Nenkova (2014); *L&N*: Li and Nenkova (2014); *B&M*: Biran and McKeown (2015); *Lin et al.*: Lin et al. (2014); *Brown-concat*: reimplementation of concatenation of Brown clusters as in Braud and Denis (2015).

- Between s_1 and s_2;
- Between s_1 and C and between s_2 and C, where C denotes two sentences immediately before s_1 (or one sentence if s_1 is the second sentence of the document).

We compute the Jaccard similarity between sentences using their nouns, verbs, adjectives and adverbs. INSTANTIATION arguments are significantly less similar than other adjacent sentence pairs (0.0335 vs. 0.0505), indicating higher differences in content. Shown in Table 3.3, both arguments of INSTANTIATION are less similar to the immediate context. While other sentence pairs follow the pattern that s_2 is much less similar to s_1's immediate context, this phenomenon is not significant for INSTANTIATION.

3.2.3 Experiments

We demonstrate the benefit of exploiting INSTANTIATION characteristics in the identification of the relation.

Settings. Following prior work that identifies the more detailed (second-level) relations in the PDTB (Biran and McKeown, 2015; Lin et al., 2014), we use sections 2-21 as training, section 23 as testing. The rest of the corpus is used for development. The task is to predict if an implicit INSTANTIATION relation holds between pairs of adjacent sentences in the same paragraph. Sentence pairs with INSTANTIATION relation constitute the positive class; all other non-explicit relations[5] constitute the negative class. We use Logistic Regression with

[5]including AltLex, EntRel and NoRel

class weights inversely proportional to the size of each class.

Features. The factors discussed previously are adopted as the *only* features in the classifier. We use the average values of s_1 and s_2 and their difference for: the number of words, difference in number of words compared to the sentence before s_1, the percentage of OOVs, gradable adjectives, POS tags and Jaccard similarity to immediate context. We use the minimum, maximum and average differences in word-pair unigram log probability, and average Jaccard similarity across sentence pairs. For Wordnet relations, we use binary features indicating the presence of a relation.

Results. To compare with our INSTANTIATION-specific classifier (*Inst. specific*), we show results from two state-of-the-art PDTB discourse parsers that identify second-level relations: Biran and McKeown (2015) (*B&M*) and Lin et al. (2014). We also compare the results with the classifier from our prior work (Li and Nenkova, 2014) (*L&N*). In that work we introduce syntactic production-stick features, which minimize the occurrence of features with zero values. Furthermore, we re-implemented Brown-cluster features (concatenation of clusters in each sentence) that have been shown to perform well in identifying INSTANTIATION's parent class EXPANSION (Braud and Denis, 2015).[6]

Table 3.4 shows the precision, recall, F-measure and balanced accuracy (average of the accuracies for the positive and negative class respectively) for each system. We show balanced accuracy rather than overall accuracy due to the highly skewed class distribution. For *Inst. specific*, we use a threshold of 0.65 for positive labels[7]. We also use *Inst. specific* along with L&N for a *soft voting* classifier, where the label is assigned to the class with larger weighted posterior probability sum from the two classifiers[8]. Both classifiers achieved at least 5% improvement of F-measure and 8%-10% improvement of balanced accuracy compared to other systems. These improvements mostly come from a dramatic improvement in recall. The improvement achieved by the voting classifier also indicate that *Inst. specific* provide complementary signals to syntactic production rules. Note that compared to Lin

[6]The dimension of clusters are tuned on the development set. As in prior work, we use clusters in Turian et al. (2010).

[7]Tuned on development set.

[8]The weights are: 0.9 for L&N and 1.0 for *Inst. specific*, tuned on development set. We also tried voting with Brown-concat but it did not outperform combining with L&N.

et al., *Inst. specific* behaves very differently in precision and recall, indicating potential for further system combination.

3.2.4 Discussion: textual entailment in INSTANTIATION

In Section 2.1, we mentioned that one way to identify INSTANTIATION is to make use of the entailment relationship $s_2 \models s_1$ (Rus et al., 2009). To study this entailment relationship, we checked how often an s_2 of an INSTANTIATION relation is automatically recognized to entail s_1. We ran the Excitement Open Platform for Recognizing Textual Entailments (Magnini et al., 2014) over the implicit INSTANTIATION arguments in the PDTB. Out of the 1,457 instances, only 20 were tagged entailment.

To understand why the recognition rate is low, we conducted an annotation task where two expert annotators are asked to mark spans in an s_1 of INSTANTIATION that is elaborated in s_2. This is illustrated in the following two examples; the INSTANTIATION arguments are in square brackets and spans being elaborated are annotated in italic:

> [1a] But industry watchers expect them [to blend the methodical marketing strategies they use for more mundane products with *the more intuitive approach typical of cosmetics companies*].
>
> [1b] [Likely changes include more emphasis on research, soaring advertising budgets and aggressive pricing].
>
> [2a] [*The pound immediately began to take a buffetting* after the resignations were announced.]
>
> [2b] [In late New York trading , sterling stood at $1.5765 , down from $1.6145 late Wednesday.]

In [1], the elaboration of [1b] corresponds to the full NP; in [2], the elaboration of [2b] corresponds to the entire clause in the beginning. Hence the entailment relationship appears to be at phrase or clause levels. It involves rich knowledge such as knowing what "the more intuitive approach typical of cosmetics companies" consists of. It also is informed by the context of events, e.g., what "late Wednesday" means for the resignation and the buffeting. This qualitatively illustrates why RTE systems do not do well on INSTANTIATION relations,

even though conceptually the two ought to be closely related; future work connecting the two may benefit both RTE and INSTANTIATION detection.

3.2.5 Conclusion

We provide the first systematic corpus analysis of the relation and show that relation-specific features can improve considerably the detection of the relation. We show that sentences involved in INSTANTIATION are set apart from other sentences by the use of gradable (subjective) adjectives, the occurrence of rare words and by different patterns in part-of-speech usage. Words across arguments of INSTANTIATION are connected through hypernym and meronym relations significantly more often than in other sentences and that they stand out in context by being significantly less similar to each other than other adjacent sentence pairs. These factors provide substantial predictive power that improves the identification of implicit INSTANTIATION relation by more than 5% F-measure.

3.3 SPECITELLER: predicting sentence specificity

Using the INSTANTIATION relation to classify sentence specificity was introduced by Louis and Nenkova (2011a). Since there is no dedicated training data available for sentence-level specificity, the two arguments of INSTANTIATION are used for training. In this setting, all sentences in an implicit INSTANTIATION arg_1 are used as general sentences and all arg_2s are labeled specific. Using a rich set of syntactic and lexical features, Louis and Nenkova's system yielded good accuracy on a test set of sentences manually labeled for specificity (Louis and Nenkova, 2012). However, the value of lexical features in predicting sentence specificity remains unclear. The experiments presented by Louis and Nenkova show that word identity features are not robust. We separately study word identity features, word properties, word embedding and clustering representations. Our experiments show that the two latter representations of lexical content are powerful and robust when trained on a large dataset using a semi-supervised approach. Furthermore these lexical features can be used as basis in exploiting unlabeled text corpora to increase the amount of available training data and significantly outperform the accuracy of prediction of the state of the art system. In this way specificity can be computed quickly so that it can become practical as a module

in realistic applications. We make our sentence specificity tool — SPECITELLER — available at `http://www.cis.upenn.edu/~nlp/software/speciteller.html`.

3.3.1 Data

To train our semi-supervised model for sentence specificity, we follow Louis and Nenkova (2011a) and use the first argument of an INSTANTIATION as an example labeled *general* and the second as an example labeled *specific*. There are 2,796 training instances in total. We then make use of unlabeled data for co-training. The unlabeled data is extracted from the Associated Press and New York Times portions of the Gigaword corpus (Graff and Cieri, 2003), as well as Wall Street Journal articles from the Penn Treebank corpus selected so that there is no overlap between them and the labeled training examples and the testing data.

For evaluation, we use the set of manual annotation of specificity by five annotators (Louis and Nenkova, 2012). Annotations cover 885 sentences from nine complete news articles from three sources—Wall Street Journal, New York Times and Associated Press. In this dataset, 54.58% of the sentences are labeled specific.

3.3.2 Light features

Shallow features. We use seven features capturing **sentence surface characteristics**. Among these, the number of words in the sentence is an important feature because on average specific sentences tend to be longer. To approximate the detection of named entities, we introduce features to track the number of numbers, capital letters and non-alphanumeric symbols in the sentence as three features, normalized by the number of words in the sentence. Symbols include punctuation so this feature captures a rudimentary aspect of syntactic complexity indicated by the presence of commas, colons and parenthesis. We also include a feature that is the average number of characters in the words that appear in the sentence, with the intuition that longer words are likely to be more specific. We also include as features the number of stop words in the sentence normalized by the total number of words, with the intuition that specific sentences will have more details, introduced in prepositional phrases containing prepositions and determiners. We use a pre-defined list of 570 stop

words provided by the NLTK package. We also include as a feature the count of the 100 words that can serve as explicit discourse connectives (Prasad et al., 2008) because explicit discourse relations within the sentence, such as elaboration or contingency, may signal that extra information is present for some of the clauses in the sentence.

We further adopt features that capture the degree to which words in the sentence **have a given property**. Louis and Nenkova (2011a) observed that general sentences tend to be more subjective. Like them, we also include the number of polar[9] and strongly subjective words (normalized by sentence length), according to the General Inquirer (Stone and Hunt, 1963) and MPQA (Wilson et al., 2009) lexicons to define two sentence features.

We also include two other dictionary features that have not been explored in prior work. We use the word norms from the MRC Psycholinguistic Database (Wilson, 1988). These are average ratings by multiple subjects of the familiarity, concreteness, imageability and meaningfulness of the word given by multiple people. We computed the cumulative ratings for words in specific and general sentences in the supervised portion of our training data. The familiarity (how familiar the word was to the subjects) and imageability (to what extent the word evoked an image according to the subjects) were significantly higher for general sentences compared to specific sentences in the "general" portion of the training data. The difference with respect to the other properties was small. So we record the average word familiarity and imageability ratings in the sentence as features.

Finally, we capture the informational value of words as approximated by their inverse document frequency (idf) weight calculated on the entire set of New York Times articles from 2006 (Sandhaus, 2008). Very common words have low idf weight and fairly rare words have high idf. We compute the minimum, maximum and average inverse document frequency values of words in each sentence, accounting for three new sentence features in this representation.

Non-sparse word representations It stands to reason that lexical features would be helpful in predicting sentence specificity, with general words characterizing general sentences. However, prior work (Louis and Nenkova, 2011a) reported that word identity rep-

[9] we refer to its usage in sentiment analysis: words that evokes something positive or negative (Wilson et al., 2009).

resentations gave very unstable results for the sentence specificity prediction task. These findings can be explained by the fact that their method is fully supervised and the training set contains fewer than three thousand sentences. In that data, only 10,235 words occur more than three times. So in new test data many sentences would have few non-zero representations other than function words because few of the content words in the training data appear in them[10]. Overall there will be only weak evidence for the association between the feature and the specificity classes. We explore two alternative representations that encode lexical information in a more general manner, tracking the occurrence of clusters of words or representing words in low dimensional dense vector space.

Brown clusters (Brown et al., 1992) are compact representations of word classes that tend to appear in adjacent positions in the training set. They were originally proposed as a way of dealing with lexical sparsity for bigram language models. In our work, we use the precomputed hierarchical clusters provided by Turian et al. (2010). The clusters are derived from the RCV1 corpus which consists of about 34 million words. Each feature in this representation corresponds to a cluster and the value of the feature is the number of occurrences in the sentence of any of the words in the cluster. The number of clusters is a parameter of the representation which we tuned with 10-fold cross validation on the labeled training data. We use 100 clusters for the results reported here.

Word embeddings are a natural product from neural network language models. In these models words are represented in low dimensional space that capture the distributional properties of words (Mikolov et al., 2013). In our experiments we use the 100-dimensional word vector representations provided by Turian et al. (2010). To represent a sentence in this space, we average the representations of the words in the sentence (Dinu and Lapata, 2010; Braud and Denis, 2015), i.e, component i of the sentence representation is equal to the average value of component i for the representations of all words in the sentence.

3.3.3 Supervised learning results

First we evaluate the feature classes introduced above in a standard supervised learning setting. We used the labeled training set to train a logistic regression classifier. We choose

[10]About 40% of our test instances have fewer than 4 content words that can be found in the labeled training data.

Features	Accuracy	Precision	Recall	F
SF	71.53	66.52	**75.12**	70.55
WP	72.43	**69.85**	69.15	69.50
Shallow (SF+WP)	**73.56**	69.44	74.63	**71.94**
BC	70.85	66.59	71.89	69.14
WE	68.25	65.24	64.43	64.83
BC+WE	71.64	70.03	65.67	67.78
Word identity	63.39	58.48	66.92	62.42

Table 3.5: Supervised learning results for sentence specificity prediction: accuracy, precision, recall and F measure on general sentences, for sentence surface features (SF), word properties (WP), combined shallow features (SF+WP), Brown clusters (BC), word embeddings (WE), and unigrams (Word identity). It is feasible to predict specificity based on cheaply computable features alone; non-sparse representations of lexical information are more suitable for the relatively small training set.

logistic regression in order to use the posterior class probability of an example being specific as a continuous measure of sentence specificity in later experiments. The models are tested on the human labeled test data.

In Table 3.5 we list the overall accuracy and the precision/recall for the general sentences achieved with each feature representation. For this test set, the majority baseline would give a 54.58% accuracy.

The class of shallow features performs reasonably well, achieving accuracy of 73.56%. This result is better than individually using surface features or word property dictionary features alone. As reported in prior work, word identity features work poorly and lead to results that are almost 10% worse than the shallow features. The non-sparse representations perform markedly better. The Brown cluster representation almost closes the gap between the lexical and shallow features with accuracy of close to 71%. Combining this with the word embedding representation leads to further small improvements. These results show that it is feasible to predict specificity based on cheaply computable features alone and that non-sparse representations of lexical information are more suitable for the relatively small training set.

3.3.4 Semi-supervised learning via co-training

In co-training, two classifiers are trained on a labeled dataset. Then they are used iteratively to classify a large number of unlabeled examples, expanding the labeled data on which they

Algorithm 1 Co-training algorithm for predicting sentence specificity

$L \leftarrow$ Labeled training examples
$U \leftarrow$ Unlabeled examples
$F_1 \leftarrow$ shallow features
$F_2 \leftarrow$ word representation features
for $i \leftarrow 0$ **to** 1 **do**
  Train classifier C_i over L using features F_i
end for
while $U \neq \emptyset$ **and** $|U|$ shrunk in the last iteration **do**
  for $j \leftarrow 0$ **to** 1 **do**
    $i \leftarrow 1 - j$
    C_i labels each example in U
    $P \leftarrow p$ examples in U most confidently labeled $+1$
    $N \leftarrow n$ examples in U most confidently labeled -1
    $K \leftarrow \{p \cup n \mid Pr_i(1|p_{\in P}) > \alpha_i, Pr_i(-1|n_{\in N}) > \alpha_i\}$
    $K' \leftarrow$ downsample(K, γ)
    $L \leftarrow L + K'$, $U \leftarrow U - K'$
    Re-train C_j over L using features F_j
  end for
end while

are re-trained. An important characteristic that ensures improved performance is that the two classifiers are independent relying on different views of the data to make decisions about the class. In our work, the shallow features and the non-sparse lexical representation provide such different views on the data, as reflected by the different precision and recall values shown in Table 3.5.

The co-training procedure for identifying general/specific sentences is detailed in Algorithm 1. It aligns with the traditional algorithm, except that we have one additional constraint as how new labeled data are added. The procedure can be viewed as a two-phase process: a supervised learning phase and a bootstrapping phase.

During the supervised learning phase, two classifiers are trained on the data from the implicit INSTANTIATION discourse relation: one with shallow features (C_0), the other with word representation features (C_1).

For the bootstrapping phase, the classifiers will take turns to label examples for each other. In each iteration, one classifier (C_i) will label each instance in the unlabeled examples. Then, at most p positive examples and n negative examples most confidently labeled are removed from the unlabeled set and added to the labeled examples. Here we set the values $p = 1000, n = 1500$. This 1:1.5 ratio is selected by tuning the accuracy of prediction on the

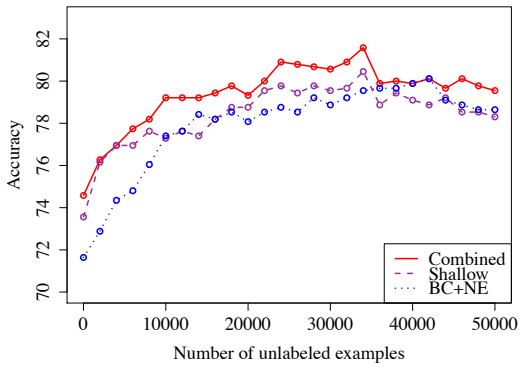


Figure 3.1: Co-training accuracies of sentence specificity prediction with increasing number of unlabeled sentences.

initial discourse training data after 30,000 new examples are added.

We impose a further constraint that the posterior probability of a new example given by C_i must be greater than a threshold α_i. The value of α_i is determined via 10-fold cross validation on the labeled training data. We choose the lowest threshold for which the prediction accuracy of the classifier on sentences with posterior probability exceeding the threshold is greater than 85%. This thresholds turned out to be 0.8 for both classifiers. To prevent a highly imbalanced data distribution, we use a procedure *downsample(*K, γ*)* in each iteration when newly labeled data is added, in which we restrict the number of samples added in the larger class to be at most $\gamma = 2$ times the size of the smaller class.

The expanded labeled examples now contain the original labeled data from discourse annotations as well as initially unlabeled instances that were confidently labeled by C_i. Now, the other classifier C_{1-i} will be re-trained using the updated labeled examples, resulting in a new classifier C'_{1-i}. C'_{1-i} will then be used to label the remaining unlabeled examples, to expand the labeled training set for C_i. The two classifiers will alternate in this fashion to label examples for each other from the unlabeled data, until no more unlabeled examples can be added.

The final prediction on the test data is decided based on the average posterior probability of labeling the sentence general from the two classifiers.

Classifier	Accuracy	Precision	Recall	F
Combined	**81.58***	**80.56**	**78.36**	**79.45**
Shallow	80.45*	79.74	76.37	78.02
BC+WE	79.55	77.42	77.61	77.52
Word identity	69.83	65.10	72.39	68.55
L&N	77.40	74.40	76.62	75.49

Table 3.6: Performance for sentence specificity prediction at the final stage of co-training, for shallow features (*shallow*), Brown clusters and word embeddings (*BC+WE*), the combination of the two (*combined*), unigrams (*word identity*) and Louis and Nenkova (2011a) (*L&N*). An asterisk (*) denotes significant improvement from L&N ($p < 0.05$, sign test). The co-training system performs the best. The performance of word identity features also improves when more data is added.

3.3.5 Semi-supervised learning results

To illustrate the effect of the larger training set obtained in co-training, we plot the classifier performance as a function of the amount of unlabeled data used for the experiments. In Figure 3.1 we show the accuracies of our semi-supervised classifiers: *i)* the dotted line represents the classifier using word representation features (brown clustering and word embeddings); *ii)* the dashed line represents the classifier using shallow features; and *iii)* the solid line represents the final *combined* classifier. The number of unlabeled data added increases from 0 to 50,000 examples, with a 2,000 step size.

The leftmost dots in Figure 3.1 correspond to accuracies without adding any unlabeled data. Initially all three classifiers gain in performance as the size of the unlabeled data grows. The performance peaks when 34,000 unlabeled examples and flattens out after this point; increasing the size of the unlabeled data is not helpful beyond this point.

At first, in each iteration, the shallow classifier almost always outperforms the word representation classifier. However, as more unlabeled examples are added, the combined classifier gains better performance as the word representation classifier becomes better and more stable. This may be due to the fact that with more data, word representations capture more and more semantic information in the sentences. Eventually, the combined classifier is much better than either one of the individual classifiers.

We thus fix our final model as the combined classifier when the benefit of adding more unlabeled data in the co-training algorithm begins to diminishes (i.e., at 34,000 unlabeled

Newly labeled general sentences	Newly labeled specific sentences
1. Edberg was troubled by inconsistent serves. 2. Demands for Moeller's freedom have been a feature of leftist demonstrations for years. 3. But in a bizarre bit of social engineering, U.S. occupation forces instructed Japanese filmmakers to begin showing on-screen kisses. 4. Although many of the world's top track and field stars are Americans, the sport has suffered from a lack of exposure and popularity in the United States.	1. Shipments fell 0.7 percent in September. 2. Indian skipper Mohammed Azharuddin won the toss and decided to bat first on a slow wicket. 3. He started this week as the second-leading rusher in the AFC with 1,096 yards, just 5 yards behind San Diego's Natrone Means. 4. The other two, Lt. Gen. Cedras and Brig. Gen. Philippe Biamby, resigned and fled into self-imposed exile in Panama two days before Aristide's U.S.-backed homecoming on Oct. 15.

Table 3.7: Examples of general and specific sentences newly labeled during the co-training procedure.

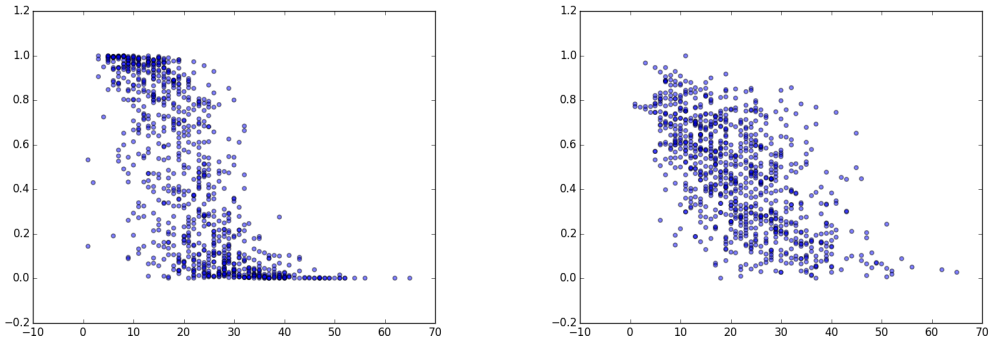
examples). In Table 3.6, we show the accuracy, precision, recall and F measure of the model on the human labeled test set. Also listed is the performance of the model proposed by Louis and Nenkova. A sign test was conducted and showed that both the combined model and the shallow model obtained via co-training is significantly better than Louis and Nenkova at 95% confidence level. Furthermore, we observe a nearly 4% increase in F measure for the combined model and higher F measure for both the shallow and word representation model after co-training. At the end of the co-training stage, with only surface features, both shallow and word representation classifiers outperform that in Louis and Nenkova.

Again, to demonstrate the effect of using word representations, we run the co-training procedure where we substitute the word representation classifier with one that is trained from word identity representations as described in the previous section. Even with more data added, lexical identity representation do not perform that well. The increased size of the training data however helps to boost the performance of the word identity representations by 1.7% in accuracy from the condition when only the original labeled data is used for training.

In Table 3.7, we show several examples of sentences from the unlabeled data that were labeled during co-training.

3.3.6 Discussion

Sentence length. During analysis we found sentence length to have a relatively high correlation (Spearman, 0.78) with the posterior probability of the sentence being specific.



(a) This work.

(b) Louis and Nenkova (2011a).

Figure 3.2: Scatter plot of sentence length (x-axis) and posterior probabilities of sentence specificity prediction (y-axis) on the test data. Our system gives more probabilities towards 0 and 1, hence the higher correlation with sentence length than Louis and Nenkova (2011a).

This is partly the result of the co-training algorithm: at each step the algorithm adds into the training set examples one of the classifiers is most confident of, pushing the probabilities more to both ends of the spectrum. We illustrate this effect in Figure 3.2 (a), compared to (b) from Louis and Nenkova (2011a) who shows smaller correlation (Spearman, 0.61). Intuitively, in (b), given a length l, there are more sentences with specificity lower than sentences with length $l + 1$, so the correlation is lower than in (a).

On the other hand, sentence length alone is not a good predictor for specificity. Using INSTANTIATION as training data, the accuracy is only 0.6936, with an F1 score of 0.6915. Instead, we find specificity to complement sentence length in applications that we have explored: identifying Chinese content-heavy sentences (Section 6.4) and identifying sentences for simplification (Section 7.1).

Specificity of generics and habituals. As described in Section 2.3, generics make distinctions between a class vs. a specific individual; habituals make distinctions between regularities vs. episodic events. Intuitively, generics and habituals are associated with more general information. To verify this hypothesis, we calculate the per-word specificity of clauses in the WikiGeneric corpus (Friedrich and Pinkal, 2015).

The WikiGenerics corpus consists of 102 texts on different topics. To obtain independent clauses automatically, these texts were segmented into 10K Elementary Discourse Units

(EDU) using SPADE (Soricut and Marcu, 2003). Each EDU is then annotated whether the clause contains a generic NP (i.e., the clause makes a statement about a class), whether the clause is generic (i.e., the clause makes a non-episodic statement about a class), and whether the clause is habitual.

We observe that predicted per-word specificity are significantly[11] lower in all three cases:

- generic clauses: 0.012 vs. 0.022, $p = 2.27e - 104$;
- clauses that contain generic noun phrases: 0.013 vs. 0.023, $p = 7.61e - 91$;
- habitual clauses 0.013 vs. 0.019, $p = 9.88e - 14$.

These results clearly show that generic expressions are associated with lower word specificity as predicted by our method.

3.3.7 Conclusion

Using the discourse relation INSTANTIATION as a seed, we presented a new model for identifying sentence specificity via co-training based on surface features that are easy and fast to compute. We make use of complementary surface features derived from the sentence and word properties as well as non-sparse word representations. The result is a lightweight model free of heavy text-preprocessing requirements that significantly outperformed the model proposed in prior work. We make the system available in our tool SPECITELLER.

3.4 Discussion: INSTANTIATION vs. SPECIFICATION

In Section 3.2, we have explored unique characteristics of sentences and sentence pairs involved in the INSTANTIATION relation. These characteristics can be exploited to substantially improve the detection of the relation. More importantly we understand why the relation is so fitting in training sentence specificity. These findings set INSTANTIATION apart from other discourse relations. However one particular relation that deserves more discussion is the SPECIFICATION relation, whose definition directly indicates changes in specificity: "arg_2 describes the situation described in arg_1 in more detail" (Prasad et al., 2007). SPECIFICATION also has more than twice the number of examples in the PDTB than

[11] Using the Wilcoxon ranksums test.

	Spec.		Inst.		
	s_1	s_2	s_1	s_2	¬ Spec.
#words	$22.0^{*\dagger}$	$26.1^{*\dagger}$	18.4	26.8	23.8
%oov	$1.02^{*\dagger}$	1.37	0.68	1.54	1.48
%gradable adj	2.65^{*}	2.03	2.96	2.22	2.22

Table 3.8: Average numbers of words, percentages of rare words and percentages of gradable adjectives for: each sentence of SPECIFICATION and INSTANTIATION, and of non-SPECIFICATION sentences. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to non-SPECIFICATION sentences. A dagger (†) denotes significant differences compared to the corresponding INSTANTIATION sentence. Characteristics in SPECIFICATION do not stand out as much as in INSTANTIATION.

INSTANTIATION, so it is tempting to use the relation to train a sentence specificity classifier. However such attempt in prior work was not successful (Louis and Nenkova, 2011a). In this section we conduct a corpus study with SPECIFICATION, similar to the one in Section 3.2.2. The study reveals that while SPECIFICATION shares some similar characteristics with INSTANTIATION, most of them distinguish the relation from others to a *lesser* degree than INSTANTIATION. The two relations are also sufficiently different in key aspects involving the use of nouns and verbs.

Table 3.8 shows the average lengths of each argument of the relation s_1 and s_2, percentages of rare words (words not in the Glove vocabulary) and gradable adjectives (e.g., *popular*, *high*) in SPECIFICATION, INSTANTIATION, and those not of SPECIFICATION. For both relations, their first sentences are on average shorter, have fewer rare words and more gradable adjectives than other sentences. However, note that in INSTANTIATION, all of these characteristics are stronger than those in SPECIFICATION: s_1s of SPECIFICATION are on average significantly longer than those in INSTANTIATION, while s_2s are significantly shorter; the fraction of rare words are significantly higher in s_1s of SPECIFICATION than in s_1s of INSTANTIATION; s_2s of SPECIFICATION also on average contain fewer gradable adjectives than other sentences, though the trend is not statistically significant.

Table 3.9 shows for the SPECIFICATION relation, the POS tags that appear significantly more often in s_1 compared to s_2 or non-SPECIFICATION sentences, and those in s_2 compared to s_1 or non-SPECIFICATION sentences. A direct comparison with INSTANTIATION is shown in Table 3.10. While we can make similar observations as with INSTANTIATION about the

$s_1 > s_2$	EX JJ NN NNP POS RB† VBZ†
$s_1 < s_2$	CC CD IN MD PRP TO VB VBN VBP
s_1 vs ¬Spec.	CC^{-} CD^{-} DT^{+} JJ^{+} MD^{-} NN^{+} NNP^{+} PRP^{-} RB^{+} TO^{-} VB^{-} VBZ^{+}
s_2 vs ¬Spec.	CC^{+} CD^{+} DT^{+} IN^{+} NNP^{-} PRP^{+} PRP\$$^{+}$ RB^{-} VBD^{+} VBZ^{-}

Table 3.9: Part-of-speech tags used significantly ($p < 0.05$) more often in the first sentences of SPECIFICATION compared to the second ($s_1 > s_2$), significantly more often in the second sentences of SPECIFICATION compared to the first ($s_1 < s_2$), and significantly more (+) or less (−) often compared to non-SPECIFICATION sentences. A dagger (†) denotes that for non-SPECIFICATION sentence pairs the difference is significant in the other direction.

s_1, Inst.>Spec.	JJ NNS RB VBG VBP
s_1, Inst.<Spec.	CD DT NN NNP TO VBD
s_2, Inst.>Spec.	NNP
s_2, Inst.<Spec.	CD DT IN MD NNS PRP VB VBN

Table 3.10: Part-of-speech tags used significantly more or less often in INSTANTIATION than in SPECIFICATION.

use of adjectives, adverbs and numbers compared to other relations, adjectives and adverbs are used significantly more often in s_1s of INSTANTIATION and numbers more often in s_1s of SPECIFICATION. Singular/mass nouns and proper nouns are more prevalent in s_1s of SPECIFICATION, even more than sentences of other relations. Moreover, we see fewer types of verbs used in s_1 compared to INSTANTIATION; s_2s of INSTANTIATION relation also on average contain significantly more proper nouns, while s_2s of SPECIFICATION contain fewer proper nouns.

For WordNet relations, only *verb group* (semantically similar verbs) appears more often among content words across the two arguments in SPECIFICATION; we also observe tendencies toward significance ($0.05 < p \leq 0.1$) for *VERB.cause* and *VERB.hyponym*. In contrast to INSTANTIATION, no noun-noun relationships are prevalent.

In terms of lexical similarity, we measure the average Jaccard similarity (using nouns, verbs, adjectives and adverbs) between the two SPECIFICATION arguments. The similarity between SPECIFICATION arguments is 0.0503, significantly less than the 0.0555 between other adjacent sentence pairs but larger than INSTANTIATION's 0.0335. In Table 3.11 we also show for both relations the average Jaccard similarity (using nouns, verbs, adjectives and adverbs) between each argument and the relation's immediate previous context (up to

	Specification			Instantiation		
	s_1	s_2	Δsim	s_1	s_2	Δsim
R	0.0330*	0.0322*	0.0008*	0.0282*	0.0275*	0.0007
¬R	0.0394	0.0342	0.0052*	0.0390	0.0358	0.0042*

Table 3.11: Average Jaccard similarity between relation sentences and their immediate context before the first sentence, for SPECIFICATION and INSTANTIATION (row 1) and sentences not of the corresponding relation (row 2). The last columns of each relation show the change between s_1 and s_2's similarities. An asterisk ($*$) denotes significant differences ($p < 0.05$) compared to sentences not of the relation. SPECIFICATION sentences are more similar to their context than INSTANTIATION.

two sentences immediately before s_1). While compared to other adjacent sentence pairs, the two arguments of the two relations are both less similar to each other, the effect is more apparent in INSTANTIATION. The drop in similarity with immediate context going from s_1 to s_2, though small, is statistically significant, while for INSTANTIATION it is not.

In sum, while SPECIFICATION and INSTANTIATION share a few characteristics among their two arguments such as length, the frequency of gradable adjectives and rare words, and similarity to each other and immediate context, the degree to which these characters stand out from other sentences are lower in SPECIFICATION. Other important characteristics, especially the use of nouns and verbs, are different between the two relations. Hence properties in INSTANTIATION are easier to capture, making it more suitable as the initial training data for sentence specificity.

In addition to the corpus study, we can also consider the difference in specificity between arguments of INSTANTIATION and SPECIFICATION. It is possible that one may provide more details in s_2 to s_1 without s_1 being a general sentence or without s_2 being a specific sentence; consider the following example:

> *An enormous turtle has succeeded where the government has failed:* **He has made speaking Filipino respectable.**

Hence one may hypothesize that the difference in specificity between arguments in SPECIFICATION is on average smaller than that in INSTANTIATION. We indeed observe this by running SPECITELLER. On average, this difference for INSTANTIATION is 0.382 (0.163 for s_1 and 0.545 for s_2), while for SPECIFICATION it is 0.154 (0.278 for s_1 and 0.432 for s_2). Not only are the s_1s significantly ($p < 0.05$) more general in INSTANTIATION and s_2s signif-

icantly more specific, the jump in specificity is significantly larger in INSTANTIATION. We note however, that this result should be better determined by a specificity measure independent of INSTANTIATION. This is not the case here since INSTANTIATION is used as the seed data for co-training in SPECITELLER. A better alternative would be to employ human judgements for specificity, where each sentence of the two relations can be assigned a rating of specificity (as we will describe in Section 4.1). We leave for future work to further explore this hypothesis.

Finally, when we used INSTANTIATION for training, all ordering and relative specificity information between s_1 and s_2 was stripped. Despite this our prediction still achieved high accuracy. Though the setting seems unintuitive, it is due to the characteristic lexical usage that not only sets the pair of sentences apart but also their individual sentences. In the next chapter we will extend this insight by exploring the connection between subsentential specificity and INSTANTIATION.

Chapter 4

Fine-grained text specificity and its connection with discourse

So far, sentences have been treated as either general or specific. Sentence specificity annotation was disconnected from specificity of finer-grained linguistic units, including noun phrase semantics, entity instantiation and generic expressions, discussed in Chapter 2. This is because using INSTANTIATION as training data inevitably results in binary labels—arg_1s as general sentences and arg_2s as specific. However, sentences often contain a mixture of general and specific content, illustrated in the following:

> Charles Hess, the director of the agency, the Project and Contracting Office, said in a telephone interview from Baghdad that the change was a natural evolution.

The named entity "Charles Hess" is specific yet terms such as "the change" and "a natural evolution" are not. Analysis of annotator disagreement by Louis and Nenkova (2012) led to the conclusion that a scale of specificity would be more appropriate and that context information should be incorporated in the annotation to resolve anaphoric and topical references that otherwise appear insufficiently specific.

We first present a pilot corpus for contextually informed sentence specificity that enables

Content in section 4.1 is published at LREC 2016 (Li et al., 2016). We especially thank Byron Wallace for his advice and feedback for Section 4.2.

the joint analysis of the *degree*, *location* and *manner* of underspecification[13] in text:

- **Degree**: the specificity of a sentence is judged on a scale rather than as a binary factor;
- **Location**: segments that lack specificity are marked within each sentence;
- **Manner**: the cause of underspecification is provided for each marked segment, along with their relationship with prior context.

An example of the annotation is shown below:

> [**sentence**] Two other former U.S. Foodservice executives, Timothy J. Lee and William F. Carter, pleaded guilty to similar charges last summer.
>
> [**question**] "similar charges": What are the similar charges? (Specified in immediate prior context)

When analyzing expressions not fully specified, we found that they fall into three cases equally frequently: anaphoric and topical references that point to some content in the expression's immediate prior context, long-distance context, or not in any prior context. Interestingly, expressions not specified in any prior content follow lexical patterns that echo those in the first argument of INSTANTIATION discussed in the previous chapter and trigger high-level reading comprehension questions such as "why" and "how".

We then design a system to predict tokens marked lacking in specificity by the annotators. Our system does not rely on costly human annotation at training time. Instead, we train a recurrent neural network with attention to predict sentence specificity; intuitively, the attention mechanism learns how much to focus on each token. We use the attention weights distributed across tokens in the sentence to rank tokens by specificity. We show promising results stronger than token specificity directly derived from sentence specificity. We also found that explicitly informing the network about approximate information related to named entities and pronouns is helpful. We end the chapter with future directions for subsentential specificity prediction.

[13] As mentioned in Chapter 2, the use of *underspecification* and *underspecified* in this thesis is different from their meaning in semantics; here we use them to refer to expressions that are lacking in specificity.

4.1 Annotation and corpus analysis

4.1.1 Goal

In the brief annotation guidelines of Louis and Nenkova (2012), the general vs. specific distinction was defined in the following way:

> "General sentences are broad statements about a topic. Specific sentences contain details and can be used to support or explain the general sentences further. In other words, general sentences create expectations in the minds of a reader who would definitely need evidence or examples from the author. Specific sentences can stand by themselves."

The aim in developing the new annotation scheme was to make more explicit what it means for a sentence to "stand on its own", while still keeping it general enough to solicit judgements from lay annotators. A sentence stands on its own if the semantic interpretation of referents can be easily disambiguated by a reader to that of the intended referent, the truth value of statements in the sentence can be determined solely based on the information in the sentence and commonly shared background knowledge, and key information about the participants and causes of an event are fully expressed in the sentence.

These three requirements cover a broad range of linguistic and semantic phenomena. For example a reference to a discourse entity may not be readily interpretable when the reference is anaphoric, by either a pronoun or definite noun phrase, when the reference is by proper name with which the reader is not familiar or the reference is generic, not referring to a specific discourse entity at all (Dahl, 1975; Reiter and Frank, 2010). Similarly gradable adjectives (Frazier et al., 2008; de Marneffe et al., 2010) like "tall", "smart" and "valuable" are interpreted according to an assumed standard. If the standard is unknown or if the writer and the reader do not share the same standard for interpreting these properties, it is impossible to verify if a sentence has the same truth value for both the writer and reader. These issues of ability to verify the truth value of a statement are directly related to Wiebe (2000)'s definition of adjective subjectivity. Sentences like "He is a publishing sensation" and "He is a valuable member of our team" are subjective because different people's definitions of what selling records are sensational or what constitutes a valuable member may differ radically. Similarly when a typical argument of a verb is missing from a

sentence (Palmer et al., 2005), the reader may have difficulty understanding the full event that is being described.

Word choice can also determine the overall specificity of a sentence, by making more explicit the manner in which an action is performed or the identity of the discourse entity, as shown by the contrast of sentence pairs like "The worker cleaned the floor" vs. "The maid swept the floor" (Stinson and Tracy, 1983; Resnik, 1995; McKinlay and Markert, 2011; Nastase et al., 2012).

The annotation we propose indirectly provides mechanisms to analyze which of the above intricate linguistic and semantic phenomena trigger the need for clarification of naive readers interested in gaining good understanding of a text. It is developed with the flexibility and intention to enable further analysis such as the classification of triggers and future refinement of annotation, to provide a practical connection between language-related applications and linguistic phenomena.

4.1.2 Methodology and corpus summary

The annotation is carried out on news articles. Each article is divided into groups of 10 consecutive sentences that the annotators would work on in one session. If the selected text was found in the middle of an article, the previous sections of the article were provided to the annotators at the start of the task for reading, but participants were not asked to annotate them.

For each sentence, the annotators rate its specificity based on a scale from 0 - 6 (0 = most specific: does not require any additional information to understand who or what is involved and what is the described event; 6 = most general). For this judgement, annotators consider each sentence independent of context.

Then they mark text segments that are underspecified, identify the cause of underspecification in the form of free text questions, and identify if these questions may be answered by information given in previous context. If the annotator chose not to ask any question, she is asked to distinguish if the sentence is most specific (i.e., no underspecified segments) or most general (i.e., the sentence conveys general information that needs no further specification). The latter types of sentences capture generics such as "Cats have four paws."

that do not refer to specific events or entities (Carlson, 2005). Agreement on annotating generic noun phrases is low (Nedoluzhko, 2013), so we adopt a higher-level annotation at the sentence level that can be done with less training and with higher agreement.

There are four types of status concerning previous context:

- **In the immediate context**: the answer to the question can be found in the two immediately preceding sentences, a distance shown to be the median length of pronoun chains in writing (Hindle, 1983). Here we use this as the effective context for pronoun resolution.
- **In some previous context**: the answer to the question can be found in the article but it is in a sentence more than two sentences before the one currently being annotated.
- **Topical**: the answer is not explicitly given in the preceding discourse but can be inferred from it.
- **None**: the answer is not explicitly or implicitly included in the preceding discourse. The author intentionally left it unspecified or it is specified in the following discourse.

Additionally, we ask the annotators to only ask questions that need to be answered in order for them to properly understand the sentence and to mark only the minimal span in the sentence which needs further specification. For example,

[**sentence**] He sued the executive of the company.

[**question**] "sued": Why did he sue? (Topical).

The annotator chose the word "sued" rather than "He sued" or "He sued the executive" because the question only relates to the act of suing.

4.1.3 Corpus statistics

The annotators are native speakers of North American English (one Canadian and two Americans). The annotation was performed on 16 articles from the New York Times dataset (Sandhaus, 2008) (13 out of the 16 are full article annotations; the annotations are all carried out from the beginning). Eight of these are politics articles and the other eight business

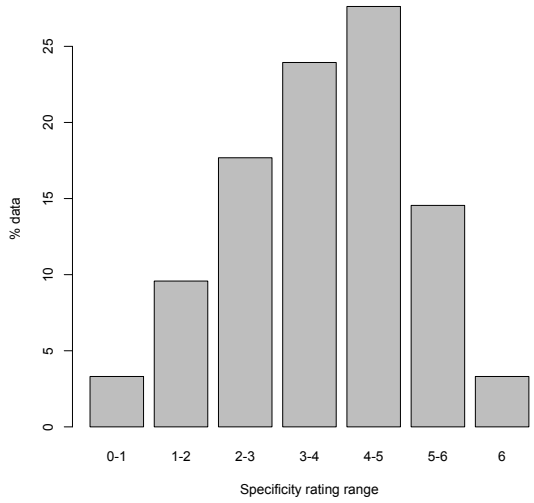


Figure 4.1: Distribution of sentence specificity ratings among the three annotators.

articles. A total of 543 sentences and 15,224 words were triple annotated by each of the annotators. The annotators generated 2,796 questions.

Sentence specificity distribution. We compute the sentence specificity score as the average from the ratings from all three annotators. Higher scores indicate more general sentences. As shown in Figure 4.1, the distribution of the ratings is roughly normal, with mean at the slightly general side. In other words most sentences are a mixture of general and specific information, confirming the need for a rating scheme rather than a binary one.

Agreement. We first compute the standard deviation of ratings among the three annotators for the sentences in the corpus. Notably, 90.4% of the standard deviation is below 1 and 64.3% below 0.5, indicating that the ratings for each sentence are close to one another.

To quantify annotator agreement we use Cronbach's α (Cronbach, 1951), which is usually interpreted as good when its values are larger than 0.8, acceptable when its values are in the 0.7–0.8 range and unacceptable when lower than 0.5. Here the annotators' α is 0.7224, which exhibits acceptable agreement. We also compare annotator agreement with the agreement one can get from random ratings. To generate the random ratings, we randomly draw a rating from the multinomial distribution given by the overall sentence specificity distribution

shown in Figure 4.1 for each sentence. This process is repeated 1,000 times and the αs are averaged. The resulting α value is 0.4886, much lower than that from the annotators and deemed unacceptable since it is lower than 0.5.

We also compute specificity rating agreement at the document level. The specificity of a document is computed as the average of the specificity ratings of the sentences in it. The correlation of document specificity scores is very high, equal to 0.98 for all three pairs of annotators.

Consensus on underspecified segments We analyze annotator agreement on identifying the location of a sentence segment that requires further specification for complete understanding of the sentence. We also tabulate the type of questions that were asked regarding the missing information. The annotators are asked to mark out the minimal text span for which she needs further specification. Each segment is associated with a free-text question and the location of the answer is given as one of *immediate context*, *previous context*, *topical*, or *none*.

The annotators asked 2,796 questions, each associated with a sentence substring (span) which the annotator identified as needing further specification. We consider three possible states for sentence substrings marked by different annotators: containment, overlap and non-overlap. Let the span of question q_i be s_i. For each question, we first check for containment among all other questions in the same sentence: $\forall j,\ s_i \in substring(s_j) \vee s_j \in substring(s_i)$. If not, we look for an overlap: $\forall j,\ s_i \cap s_j \neq \emptyset$. If neither containment nor overlap is found, we assign the "non-overlap" state to the question.

The percentage of questions with each state is: non-overlap: 0.3%; overlap: 29.8%; containment: 69.9%. It confirms that when an annotator identifies an underspecified segment, it is 99.7% likely that a part or all of the segment is also identified as underspecified by at least one other annotator. This means that the readers reach a natural consensus as of which part of the sentence needs further detail. Furthermore, the majority (69.6%) of these segments fully overlap with another.

We also calculate the percentage of underspecified tokens that are marked by one (60%), two (29.2%) or all three annotators (13.8%). Despite the high overlap of *segments* demonstrated above, there is a high percentage of tokens marked by only one annotator. This

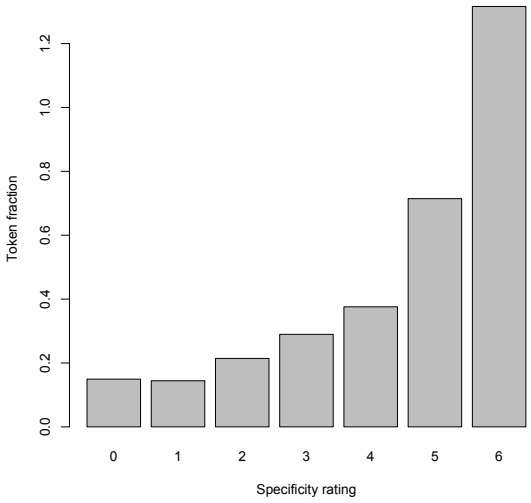


Figure 4.2: Average fraction of tokens marked as underspecified vs. average sentence specificity ratings.

shows that despite the minimality principle, identifying underspecified *tokens* of high agreement requires additional filtering.

Sub-sentential vs. sentential specificity Since the annotators are asked to give specificity ratings and ask questions independently, we can now compare number of underspecified segments at the sub-sentence level with the specificity of the overall sentence. For the former, we calculate in each sentence, the percentage of tokens marked as underspecified by at least one annotator. If an annotator did not ask a question and marked the reason to be that the sentence is too general, then a count 1 is added to all tokens in the sentence. Figure 4.2 shows that the more general the sentence was judged to be, the larger its portion of underspecified tokens.

4.1.4 Discourse analysis of underspecification

Specificity and content density The annotation of articles using a scale of specificity score allows us to study the connection between text specificity and content density. The latter, described in Yang and Nenkova (2014), represents how much the text is factual and how well the content is expressed in a "direct, succinct manner". Specifically, our

articles overlap with those annotated in Yang and Nenkova (2014), so we compare the content density scores of lead paragraphs annotated by Yang and Nenkova (2014) with their specificity. For each lead paragraph, its content density is a real-valued score assigned by two annotators (here we take the average). A larger value indicates more density. Its specificity score is obtained by averaging the sentence specificity ratings (for each sentence its specificity rating is averaged among annotators). We observe a significant ($p \leq 0.05$) Spearman correlation of -0.51, indicating that content-density on the paragraph level is positively associated with its sentences being more specific.

Entity co-reference We analyzed the connection between co-reference resolution and context-dependent underspecification (questions about content missing in the sentence but found in preceding context and necessary for full comprehension of the sentence). It is reasonable to assume that all questions resolved in the previous context involved anaphoric references to previously mentioned entities. Yet, of the underspecified segments annotated as having the missing details in the local context (i.e., two sentences above), only 34.4% contain an entity that is resolved by automatic coreference resolution[14]. For non-local previous context, this number is 26% (21.5% for all segments). This confirms that our corpus captures coherence patterns beyond noun phrase anaphora resolution problems; for instance, the example below illustrates an event coreference:

> After a contest that had pitted domestic pride against global politics, the Pentagon yesterday chose an international team, headed by Lockheed Martin, to build the next fleet of presidential helicopters over Sikorsky Aircraft, which had positioned itself as the "all-American" choice. *In selecting Lockheed, which will receive $ 1.7 billion initially to begin the program, the Pentagon signaled a new openness to foreign partners on sensitive military tasks.*
>
> **Question:** "selecting" — What were they selected for? (immediate context)

Underspecified tokens and context To support understanding of document level coherence, we link each sub-sentential underspecified text segment with the running discourse by annotating the location of answers to the question associated with each segment. The

[14]We used the Berkeley Entity Resolution System (Durrett and Klein, 2014).

Interrogative	All	%Immediate	%Previous	%None
what	1388	**36.6**	36.5	20.0
who	419	**52.7**	31.7	11.5
how	332	4.5	10.2	**76.2**
why	317	10.4	24.3	**50.5**
which	242	**40.9**	35.2	21.1
where	66	36.4	**37.9**	22.7
when	24	20.8	12.5	**62.5**

Table 4.1: Number of question interrogatives used by the three annotators and percentages of the context status associated with each question. Largest values in each row are bolded. "How", "why" and "when" questions have stronger association with answers not present in prior context.

percentages of questions whose answers can be found in the four cases: in immediate context (32.47%), in previous context (30.87%), is topically related (7.37%), or not in any prior context (28.97%). The context status of underspecification is remarkably even in the none, immediate and previous context cases, with a small portion being topical.

The type of question—signaled by the question word—gives insight to what type of information a reader is seeking to understand the sentence. The context status of the question gives information for each segment where it can be specified in the running discourse. In Table 4.1, we tabulate the number of interrogatives found in the questions along with the context status associated with each interrogative, sorted by the frequency of the interrogative. The most frequent interrogative is "what", followed by "who", "how", "why" and "which"; "where" and "when" questions are not often raised by the annotators[15]. These question words also distribute very differently in each context status; for example, most of the underspecification leading to "what", "who", "which" and "where" questions can be resolved in prior context, but "how", "why" and "when" questions are raised mostly when the lack of specificity cannot be resolved in prior context.

To study the characteristics of the tokens associated with one thirds of the lack of specificity that cannot be resolved in prior context, in Table 4.2, we lay out the percentage of universal part-of-speech tags (Petrov et al., 2012) of tokens in underspecified segments their percentage associated with the following: fully specified, resolved in immediate context, in

[15]Note that interrogatives and question types do not have a one-to-one mapping. For example, not all "what" questions are entity-centric. We found 186 of these questions that are potentially causal questions, with presence of the words *happen, reason, for, cause, mean, entail, purpose*. We leave for future work a detailed classification of question types.

POS tag	Specified	Immediate	Previous	None
ADJ	69.0	5.7	6.3	19.8
ADP	93.7	1.8	1.6	2.8
ADV	72.6	7.9	5.2	14.9
CONJ	93.1	2.1	1.9	3.7
DET	75.8	9.1	5.6	9.8
the	68.5	12.0	14.4	6.8
NOUN	71.32	7.8	12.7	10.3
NUM	88.29	5.1	4.8	3.2
PRON	67.3	21.8	12.2	1.9
PRT	90.4	2.1	2.1	4.5
VERB	82.5	3.6	4.6	9.4

Table 4.2: Percentages of part of speech tags that are not highlighted (*specified*) and those that are marked as underspecified with associated context status (*immediate, previous, none*). Most of the underspecification are from content words; among them, adjectives, adverbs and verbs have stronger association with answers not present in prior context.

previous context and no context. We also separated the definite determiner "the" from the main determiner category to distinguish between definite and indefinite references. Each token is counted once if marked by multiple annotators. These numbers clearly show that most of the underspecification comes from content words. Among them, most of the lack of specificity of pronouns and determiners can be resolved in prior context. The definite expression "the" behaves differently from indefinites; it is one of the most often marked POS tags (and most of them can be resolved in context), while other determiners are marked much less often, with a large portion that cannot be resolved in context. On the other hand, the lack of specificity from adjectives, adverbs and verbs more often cannot be resolved in context. This may explain some of the findings concerning INSTANTIATION discussed in Section 3.2: the first arguments of INSTANTIATION also contain on average more adjectives, adverbs and certain classes of verbs.

This information when combined with interrogative breakdown in Table 4.1 illustrates that underspecified content, when not elaborated before, is more likely to be non-entities and triggers high level comprehension questions.

Context status and sentence number According to the entropy constancy theory (Genzel and Charniak, 2002; Jaeger and Levy, 2007; Frank and Jaeger, 2008; Jaeger, 2010), writers refer to previously mentioned terms without fully specifying them to avoid redun-

dancy. Hence we hypothesize that the context status of underspecified segments changes with sentence number. Indeed, the position of a sentence within the document is:

- negatively correlated with the number of *not-in-context* questions ($r = -0.14, p = 6e-4$);
- positively correlated with the number of *in-context* questions and relative sentence position ($r = 0.20, p = 1e-6$);
- not correlated with the number of *immediate* questions ($r = 0.05, p = 0.22$).

So the further along the reader gets into an article, the less likely they are to ask about concepts not previously established in the context. Instead, the reason for the lack of specificity is more and more likely to be that the reader is now an "expert" in the topic and the context is assumed. This confirms our hypothesis and indicates that the later a sentence appears in a document, the harder it is to process without context.

4.1.5 Conclusion

In this section, we present an annotation method and a corpus for context-informed sentence specificity. Our methodology enables joint annotation on sentential specificity, subsentential underspecified expressions and their context dependency. We annotate the type of underspecification using high level questions generated by the annotators. We showed that the annotators reached good agreement on sentence and document level specificity and they have high consensus as which text segments within the sentence are underspecified. We have released our dataset at `http://www.cis.upenn.edu/~nlp/corpora/lrec16spec.html`.

4.2 Predicting subsentential specificity

In SPECITELLER (Section 3.3), we bootstrapped distributed word representations in INSTANTIATION to predict sentence level specificity. With the above corpus annotation and analysis, we further understand that:

- Sentences are usually a mixture of general and specific content;
- Sentence level specificity reflects the specificity of individual expressions within the sentence (Figure 4.2);

- An expression lacking in specificity often triggers a reader to ask questions for clarification or further understanding.

Knowing which expressions in the sentence are underspecified and are likely to give rise to questions can be useful in a range of applications. For example, in argumentation analysis, these expressions can be an informative indicator of the quality and importance of an argument (Swanson et al., 2015). In writing quality assessment and assistance, they can be useful in identifying vague arguments that need elaboration, for both human and machine generated text. When extending into human-robot conversation, this can be an additional signal to help to tailor machine response to the appropriate specificity level (Li et al., 2017a). Since we are interested in coherence analysis among different groups of audiences, subsentential specificity also makes it possible to automatically pinpoint specific elements in a document that can lead to differences in specificity perception and discourse understanding.

We present the first work that predicts *tokens* within a sentence that are lacking in specificity and hence are more likely than others to trigger questions (henceforth "question tokens"). Since there is no large dataset currently available to train systems to predict the specificity of subsentential units, we start our experiments using existing resources for sentence specificity and hope to derive insights about possible annotation efforts in the future. We train a neural network model for sentence-level prediction with token-level attention mechanism, which we use to extract words that are most general or specific within a sentence. We hypothesize that words the network pays the most attention to in a predicted general sentence are more likely to be underspecified, while the opposite is true for sentences that are predicted specific.

We contrast our approach to one that derives token specificity using log-odds ratios from SPECITELLER's training data and predictions of sentence specificity. The attention-based models show superior performance predicting question tokens. Although the overall performance of the task is low, there is a clear difference between the performance of the two. The most substantial improvements are achieved when predicting question tokens whose answers cannot be found in prior context (henceforth "not-in-context" tokens). As we have pointed out previously, these tokens are more likely to trigger high-level text understanding

questions such as "how" and "why". Furthermore, we show that our model also outperforms SPECITELLER on sentence-level specificity prediction on our refined dataset in the previous section (Section 4.1).

We further draw from insights from our prior work regarding pronouns and named entities. Named entities represent specific information and are approximated and included as features in SPECITELLER, while pronouns unresolved in the sentence indicate a lack of specificity. We compare different models where pronouns and named entities are approximated as additional special symbols from the input. We found that explicitly considering potential named entities is most helpful predicting question tokens, especially not-in-context tokens. On the other hand, explicitly marking potentially unresolved pronouns helps sentence level specificity prediction.

4.2.1 Related work

Our work is most closely related to research that seeks explanations for the prediction made by a neural model via an attention mechanism. The attention mechanism was first introduced in machine translation for word alignments (Bahdanau et al., 2015). It has since been successfully applied to numerous NLP tasks such as question answering (Sukhbaatar et al., 2015; Hermann et al., 2015; Kumar et al., 2016), sentiment analysis (Wang et al., 2016), text classification (Yang et al., 2016; Zhang et al., 2016) and discourse relation recognition (Liu and Li, 2016). Attention weights have been shown through visualizations to provide qualitative explanations behind the prediction (Liu and Li, 2016; Yang et al., 2016). In this work, we explicitly make use of the attention weights to predict token specificity.

Also related is a line of work that uses explicit rationales to aid predictions (Zaidan et al., 2007; Marshall et al., 2016; Zhang et al., 2016). However these methods rely on gold-standard annotations of rationales. Our method does not rely on such annotation during training. Lei et al. (2016) extracts rationale text segments such that they are sufficient to replace the original input text for the final prediction. On the contrary, we seek underspecified tokens even when the sentence is predicted specific. Our model infers a distribution of weights over all tokens, which can be used for prediction in both general and specific directions.

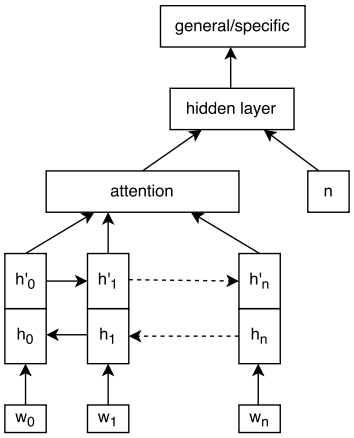


Figure 4.3: Architecture of the attention network for sentence specificity prediction.

4.2.2 Attention network for token specificity

The basis of our attention network is depicted in Figure 4.3. Its components are detailed below.

Sentence encoder. Given a sentence s, each word w_t within the sentence is first embedded into a vector x_t. Here we use Glove (Pennington et al., 2014) for word embeddings, the same as in SPECITELLER. We do not tune the embeddings.

We then use a bidirectional Long-Short Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) to encode sentences. An LSTM is a type of recurrent network with memory cells, enabling it to capture long-term dependencies. With a bidirectional LSTM, the combined hidden state h_t for each word w_t incorporates information from both before the word (i.e., w_0 to w_t) and after the word (i.e., w_t to w_n) for a sentence of n words: $h_t = [\vec{h}_t, \overleftarrow{h}_t]$.

Attention. We use the attention mechanism to impose a distribution of weights over all tokens in a sentence, such that those most informative for the prediction are assigned higher

weights (Yang et al., 2016):

$$u_t = \tanh(W_u h_t + b_u) \tag{4.1}$$

$$a_t = \text{softmax}(u_t^T w_a) \tag{4.2}$$

$$h_s = \sum_t a_t h_t \tag{4.3}$$

where W_u, b_u, w are model parameters. The representation for a sentence s is a weighted sum of hidden states h_t for all tokens; the attention weights are hence obtained by a_t.

Sentence length and prediction. Besides words in a sentence, sentence specificity is also influenced by its length. This is discussed in Section 3.3.6. To reduce the effect of length on the attention weights, we explicitly inform the network of the length of the sentence. For a sentence of length n, we feed into the final hidden layer the sentence length n along with the hidden sentence representation s, before applying the sigmoid function for binary classification:

$$y = \text{sigmoid}(W_s[h_s, n] + b_s) \tag{4.4}$$

Given the attention weights for each token and the number of tokens to output as k, we predict question tokens in the following way:

- Sentence predicted general: k tokens with the highest weights;
- Sentence predicted specific: k tokens with the lowest weights (since the highest weights are assigned to those that are most specific).

4.2.3 Named entities and pronouns

In SPECITELLER, the fraction of words with capital letters is used as features to approximate named entities. This was done to capture the intuition that named entities are associated with specific information. Indeed, the second arguments of INSTANTIATION contain significantly more of them than in other sentences. However, their effect on subsentential specificity is less clear. Named entities are specific if the readers know who/what the entity is and unspecified otherwise (Siddharthan et al., 2011). To see its effect on both sentence and

word specificity prediction, we strip capitalization but introduce a special symbol `NAMEENT`, that concatenates with each capitalized word not at the beginning of the original sentence.

Pronouns occur significantly more often in the second argument of INSTANTIATION relations than in other sentences (c.f. Table 3.2), so they may also indicate that the sentence is specific; however a pronoun not resolved within the same sentence clearly needs clarification for the sentence to be fully specified. We make a simple approximation that third person personal pronouns at the beginning of a sentence are not resolved within the sentence. We introduce a second symbol `UNRPRON`, to concatenate with each of those.

4.2.4 Systems and settings.

For training, we use all sentences used to train SPECITELLER. This includes sentences in a implicit INSTANTIATION relation from the PDTB, as well as the 34K unlabeled sentences bootstrapped by SPECITELLER's co-training algorithm.

We train four attention networks:

- **attn**: attention network without special symbols for pronouns or named entities;
- **attn-ne**: attention network with special symbol for named entities;
- **attn-pron**: attention network with special symbol for pronouns;
- **attn-ne-pron**: attention network with both special symbols.

For preprocessing, we strip all capitalization and replace numbers with a special symbol `NUM` to reduce sparsity (we additionally run a system **attn-withnum** without processing numbers to illustrate the effect). We map each word to its 100-dimension Glove embedding (Pennington et al., 2014). This setting is consistent with SPECITELLER. Out-of-vocabulary words are assigned to random, 100-dimension vectors. Special symbols and parameters are randomly initialized. The LSTM hidden state dimension is 128 (hence the bidirectional LSTM dimension is 256), tuned with validation data (validation split is 20% of the training data). The training objective is the cross entropy loss; we use stochastic gradient descent for training. A dropout layer is applied before the final layer to prevent overfitting. The dropout rate is 20%, tuned with validation data.

Classifier	Accuracy	Precision	Recall	F
attn-withnum	73.79	73.22	66.67	69.79
attn	77.51	74	77.86	75.88
attn-ne	**79.32**	**75.64**	80.35	77.93
attn-pron	77.06	73.52	77.36	75.39
attn-ne-pron	79.21	75.00	**81.34**	**78.04**
SPECITELLER	**81.58**	**80.56**	**78.36**	**79.45**

Table 4.3: Performances for sentence specificity prediction on the same data as in Table 3.6. *Attn-withnum*: attention network without processing numbers; *attn*: with special symbol for numbers (default); *attn-ne*: with special symbol for named entities; *attn-pron*: with special symbol for unresolved pronouns; *attn-ne-pron*: with both named entity and unresolved pronoun symbols. Precision and recall are on general sentences. First and second best values for each measure are bolded.

4.2.5 Sentence specificity prediction.

Before diving into our main task, we first compare the models in terms of sentence specificity prediction.

Table 4.3 shows performance on the same evaluation data as SPECITELLER, annotated by Louis and Nenkova (2011a). Precision and recall are calculated on general sentences. The attention network without mapping numbers to the special symbol `NUM` (*attn-withnum*) performs the worst, indicating that by normalizing numbers—hence reducing vocabulary sparsity—is very helpful in this task. We hence exclude *attn-withnum* from further analysis. Neural models in general give higher recall and lower precision. However, they do not outperform SPECITELLER in terms of F measure. The most notable improvements in all measures are achieved by *attn-ne*, i.e., the one explicitly informs the system about possible named entities. Marking sentence initial pronouns does not bring improvements by itself, but combined with named entity markers, it brings further improvement on recall.

We now evaluate sentence specificity with our dataset described in Section 4.1. The sentences are rated from 0 (most specific) to 6 (most general). Here we take all sentences whose average ratings are strictly below 3 to be specific sentences and all sentences whose average ratings are strictly above 3 to be general sentences[16]. There are 356 general sentences and 217 specific sentences. There are more general sentences than specific in this dataset, while in Louis and Nenkova (2011a), 55% of the sentences are specific. This is

[16]For this particular task, we do not evaluate on sentences whose average ratings are exactly 3.

Classifier	Accuracy	Precision	Recall	F
attn	71.38	91.74	59.27	72.01
attn-ne	65.62	**97.60**	45.79	62.33
attn-pron	**72.95**	93.89	**60.39**	**73.50**
attn-ne-pron	68.76	97.33	51.12	67.03
SPECITELLER	71.38	91.04	57.58	71.43

Table 4.4: Performance for sentence specificity prediction on our annotated data described in Section 4.1. *Attn*: attention network with special symbol for numbers (default); *attn-ne*: with special symbol for named entities; *attn-pron*: with special symbol for unresolved pronouns; *attn-ne-pron*: with both named entity and unresolved pronoun symbols. Precision and recall are on general sentences.

likely the result of our refined annotation guidelines, where definition for the most general and specific sentences are clarified and each underspecified expression is rationalized by the annotators.

System performances are tabulated in Table 4.4. Due to the differences in the proportion of general sentences, recall values are lower than precision. The attention models *attn* and *attn-pron* both outperform SPECITELLER in precision, recall and F measure. These two models are good at maintaining an already good precision while improving recall, making them more reliable and better in identifying general sentences. The best performance is achieved by *attn-pron*. Providing pronoun information explicitly, even though only approximately, is helpful for the models to learn to separate pronouns that are specified within the sentence vs. those that are specified elsewhere. Finally, explicitly marking named entities gives the network extra boost in precision, reaching an impressive 97%, however at much of a cost in recall. In other words, doing so makes the model to overly associate named entities with the sentence being specific.

4.2.6 Predicting question tokens

We now discuss our main task: predicting tokens that are lacking in specificity, i.e., "question tokens". In Table 4.5, we show the average and standard deviation of the number of question tokens per sentence, for all question tokens and not-in-context tokens (*N context*).

In this work, we do not solve the task of predicting how many tokens are asked about, hence we adopt two evaluation methods. First, we assume the number of question tokens are

	mean	std.dev
All	5.6	3.6
N context	2.3	2.8

Table 4.5: Average and standard deviation of the number of question tokens per sentence, for all question tokens and question tokens that are not-in-context.

	All		N context	
	Accuracy	F	Accuracy	F
attn	0.7274	0.3116	0.8305	0.2067
attn-ne	**0.7469**	**0.3339**	**0.8495**	**0.2651**
attn-pron	0.7353	0.3311	0.8338	0.2229
attn-ne-pron	0.7378	0.3000	0.8343	0.1754
speciteller	0.7310	0.3298	0.8262	0.1985
random	0.7153	0.2927	0.8249	0.1924

Table 4.6: Accuracy and F measure for token specificity prediction, when the number of question tokens is known. Column *all* shows all question tokens; column *N context* shows question tokens that are not-in-context.

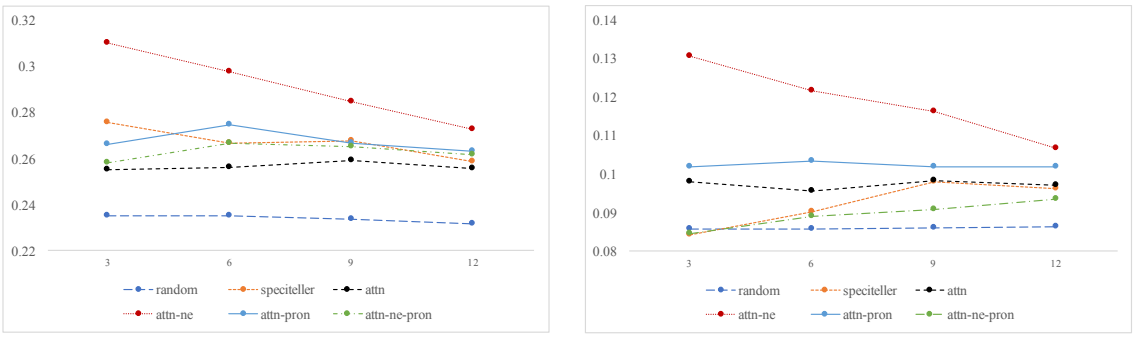
known, and look at the F measure achieved by each system. Second, we look at precision@k, a metric mostly used in information retrieval that considers the number of tokens correctly marked as question tokens in the top k tokens ranked by each model.

Baselines. For benchmarking, we consider two baselines: *random* and *speciteller*. The *random* baseline selects random tokens from the sentence given the number of tokens to select.

We also compare against SPECITELLER. Since SPECITELLER uses a combination of features other than tokens in a sentence, we derive token specificity from its training data (with 34K sentences used in the final step of co-training). For each word w, we compute its log odds ratio to measure its tendency to appear in a general sentence vs. specific:

$$LOR(w) = \log(\frac{P(w|\text{sentence=general})}{P(w|\text{sentence=specific})}) \tag{4.5}$$

Then, given the number of question tokens k, the *speciteller* baseline outputs the top k words that are most general within the sentence.



(a) All question tokens.

(b) Not-in-context question tokens.

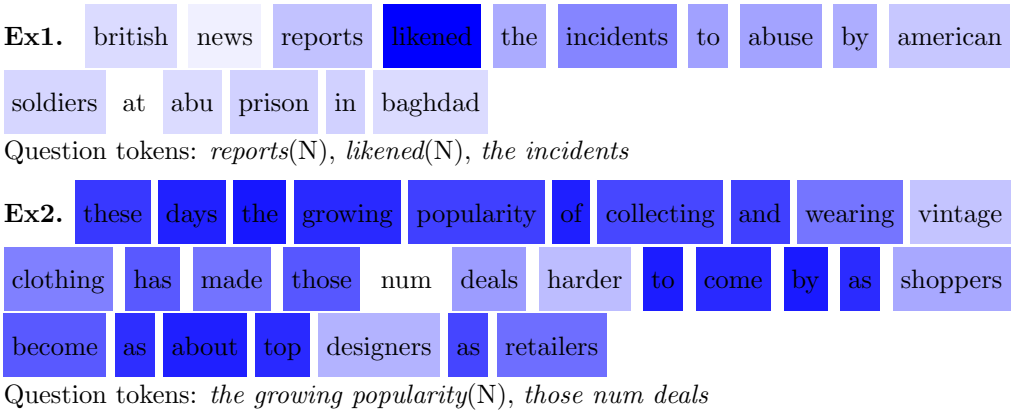
Figure 4.4: Precision at 3, 6, 9, 12 tokens for token specificity prediction. X-axis: number of tokens; y-axis: precision.

Results. First we conduct an "oracle" evaluation where we assume that the number of question tokens is known for each sentence. Table 4.6 shows accuracies and F measures for each system, for all question tokens and not-in-context tokens. The best system in both cases is *attn-ne*. While *speciteller* performed much better than *random* for all question tokens, it performs only comparable to *random* for the not-in-context case. On the other hand, the attention models are notably better in predicting not-in-context question tokens. Among the attention models, explicitly marking named entities help the most; on the other hand, explicitly marking both named entities and pronouns leads to a drop in performance, despite that they are each individually helpful.

We also conduct an evaluation where the number of question tokens is not known. Since we are not predicting how many tokens trigger questions, we consider precision values among the top 3, 6, 9 and 12 tokens for each system. These are shown in Figure 4.4. The trends are similar to the oracle evaluation. When predicting all question tokens, the attention models perform similarly to *speciteller* and better than the random baseline. When predicting only not-in-context tokens, *attn*, *attn-ne* and *attn-pron* all outperform *speciteller* and *random*. Interestingly, while *speciteller* and *attn-ne-pron* performs similarly to *random* when the number of tokens to retrieve is small (3 tokens), the precision improves when more tokens are retrieved.

Analysis. We first visualize several examples to illustrate attention distributions for the *attn* model. In each example, we show the processed text with background color whose

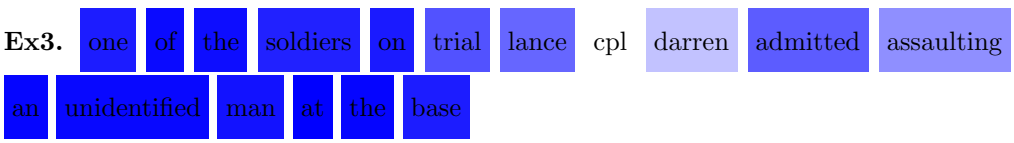
darkness reflects the token's likeliness of being a question token (darker means higher likeliness). The weights are normalized with min/max normalization; if a sentence is predicted specific, then for each weight w, we flip its value to be $1 - w$. We also show the question tokens, with a marking *(N)* for not-in-context ones:



In the first example, the neural model correctly identifies "likened" to be a question token; it is also a not-in-context token. The next token predicted is "incidents", also questioned. However the not-in-context token "reports" was not assigned the highest weight.

In Example 2, the sentence is predicted specific, so the weights are flipped. The highest weights are mostly assigned to the beginning portion of the sentence, which correctly corresponds to the not-in-context segment "the growing popularity". The model assigns "num" the least weight for questioning, since numbers are associated with very specific information in general. The model fails to assign "deals" with a high weight, even though the phrase "those num deals" is marked to be underspecified.

The *attn* model also seems to be able to capture named entities. In example 1, named entities "british", "american", "abu" and "baghdad" are assigned low weights even though they are not capitalized, indicating some capacity of the model to learn that these words are usually specific without knowing them being named entities. In the following example, the sentence is predicted specific; "CPL Darren" is a named entity and is weighted highest for the prediction (i.e., with the lightest background color for question tokens):



However, difficulty comes when named entities consist of common words. In example 4, "the way we were" is such an entity but each token is assigned a high weight for question tokens. In example 5, by adding `NAMEENT`, the system *attn-ne* is informed that it is likely a named entity, hence is not likely to trigger a question:

Ex4. [**attn**] the business of vintage clothing has changed considerably since the late num 's when small shops popped up to accommodate young buyers looking for fashion alternatives said doris raymond the owner of the way we wore a store in los angeles

Ex5. [**attn-ne**] the business of vintage clothing has changed considerably since the late num 's when small shops popped up to accommodate young buyers looking for fashion alternatives said doris NAMEENT raymond NAMEENT the owner of the NAMEENT way NAMEENT we NAMEENT wore NAMEENT a store in los NAMEENT angeles NAMEENT

Yet marking named entities in this way lead the model to overly predict sentences to be specific, as shown in Table 4.4 for sentence specificity prediction. For instance,

> Both men are directors of TV Azteca.

This is a sentence annotated as general by our annotators. The token "Azteca" is out of the training vocabulary; with the absence of this word, the *attn* model correctly predicted that it is general. However, for *attn-ne*, the sentence becomes:

> both men are directors of tv NAMEENT NAMEENT .[17]

The system then predicts that it is specific. In another example, the *attn-ne* system is able to de-emphasize the named entity that *attn* did not, but doing so drops the confidence for prediction quite a bit:

Ex6. [**attn, General, 0.99**] that 'd be good " mr wallace said

Ex7. [**attn-ne, General, 0.54**] that 'd be good " mr NAMEENT wallace NAMEENT said

[17] The token "Azteca" is out-of-vocabulary, hence it does not have a weight.

4.2.7 Conclusion

We design a system to predict tokens in a sentence that are lacking in specificity and thus are more likely to trigger questions from the reader. The core of our approach is an attention network, which we use to predict sentence level specificity while leveraging the attention weights to rank question tokens. We further explore approximate representations for named entities and pronouns that are not resolved within its sentence. We found that our attention networks, especially when informed of potential named entities, outperforms SPECITELLER in question token prediction. Finally, informing the system about potential unresolved pronouns helps with sentence specificity prediction.

4.3 Discussion and future work

We first develop an annotation scheme for sentence specificity that is suitable for a complex discourse phenomenon. This scheme not only refines the annotation for sentence specificity, but also contains information as to what expressions in the sentence are underspecified and the types of questions they trigger in readers.

We then predict these question tokens, using an attention mechanism within a sentence specificity classifier. While we have some success with this first attempt, merely using the attention weights is just a start; there are a lot of aspects of this rich task that we have not captured, for example, context and phrases. We lay out several practical extensions to this work.

Our current systems are trained on the same data as in SPECITELLER. One obvious future direction is to train on a much larger, but noisy dataset automatically labeled by SPECITELLER. Doing so not only can let the system handle a much larger vocabulary, but also makes it possible to handle more parameters. For example, in this work we take the word embeddings as-is; with more data we will be able to tune the embeddings with specificity. We will also be able to make use of more powerful (and hence more complex) neural architectures.

One of these more powerful architectures we would like to make use of is repeated attention. From our analysis, we notice that often the weights for the same sentence,

especially for words close together, have very similar weights. Repeated attention networks have been shown to "sharpen" these attentions, so the weights are pulled away from each other (Kumar et al., 2016; Liu and Li, 2016).

We would also like to extend this work beyond tokens. Often questions are based on a segment of text, e.g., "the growing popularity", "the incidents", etc. Although the current attention networks seem to capture this somewhat (e.g., tokens in "the business of" from examples 4 and 5 are assigned almost the same weights), this cannot be captured or made use of. A structured model (Kim et al., 2017) on the other hand will be able to capture specificity on the phrase level.

We would also like to explore automatic detection of the source of the lack of specificity along with how it is associated with context, for example, subjectivity, generic expressions, missing verb argument, entity instantiations and anaphora. Finally, we have not tackled the problem of how many question tokens to retrieve within a sentence. We will explore this in future work.

Chapter 5

Coherence preferences: cross-lingual

Languages differ in how information is organized into sentences. We study two types of such differences in the context of Chinese to English translation: (1) how much content is conventionally acceptable in a sentence, and (2) the use of discourse devices. We show that without properly handling these differences, the intelligibility of translated text can be problematic, especially for machine translation systems, which usually translates a single sentence in one language into a single sentence in another.

We start with two languages: Chinese and Arabic, and we present a study of aspects of discourse structure that significantly impact the quality of machine translation. Our analysis is based on manual evaluations of translations of news from Chinese and Arabic to English. We find that for Chinese, the need to employ multiple explicit discourse connectives (*because*, *but*, etc.), as well as the presence of a CONTINGENCY relation, are related to lower translation quality. The mismatches between discourse expressions across both languages also significantly impact translation quality. Furthermore, we find that there is a particularly strong mismatch in the notion of what constitutes a sentence in Chinese and English, which occurs often and is associated with significant degradation in translation quality. Although

Content in Section 5.1 is published at ACL 2014 (Li et al., 2014). Content in Section 5.2 is published at EMNLP 2015 (Li and Nenkova, 2015a).

in Arabic this type of mismatch also exists, it does not significantly impact translation quality.

We then further study this mismatch in the notion of a sentence, focusing on Chinese. In Chinese to English translation, information conveyed by some sentences would be more easily understood by a reader if they were expressed in multiple English sentences. We call such sentences *content heavy*: these are possibly grammatical but difficult to comprehend, cumbersome sentences. We develop methods to identify sentences in Chinese for which English speakers would prefer translations consisting of more than one sentence. We base our analysis and definitions on evidence from multiple human translations and reader preferences on flow and understandability. We show that machine translation quality when translating content heavy sentences is markedly worse than overall quality and that this type of sentence is fairly common in Chinese news. We demonstrate that sentence length and punctuation usage in Chinese are not sufficient clues for accurately detecting heavy sentences and present a richer classification model that accurately identifies these sentences.

5.1 Assessing the Discourse Factors that Influence the Quality of Machine Translation

In this study we examine how the use of discourse devices to organize information in a sentence — and the mismatch in their usage across languages — influence machine translation (MT) quality. The goal is to identify discourse processing tasks with high potential for improving translation systems.

Historically MT researchers have focused their attention on the mismatch of linear realization of syntactic arguments (Galley et al., 2004; Collins et al., 2005), lexico-morphological mismatch (Minkov et al., 2007; Habash and Sadat, 2006) and word polysemy (Carpuat and Wu, 2007; Chan et al., 2007). Discourse structure has largely been considered irrelevant to MT with very few studies (Marcu et al., 2000; Tu et al., 2013; Guzmán et al., 2014), mostly due to the assumption that discourse analysis is needed to interpret multi-sentential text while statistical MT systems are trained to translate a single sentence in one language into a single sentence in another.

However, discourse devices are at play in the organization of information into complex sentences. The mere definition of sentence may differ across languages. Chinese for example is anecdotally known to allow for very long sentences which at times require the use of multiple English sentences to express the same content and preserve grammaticality. Similarly discourse connectives like *because, but, since* and *while* often relate information expressed in simple sentential clauses. There can also be possible complications in translating connectives, for example, explicit discourse connectives may be translated into implicit discourse relations or translated in morphology rather than lexical items (Meyer and Webber, 2013; Meyer and Poláková, 2013).

In our work, we quantify the relationship between information organization, discourse devices, and translation quality.

5.1.1 Data and experiment settings

We examine the quality of translations to English from Chinese and Arabic using Human-targeted Translation Edit Rates (HTER) (Snover et al., 2006), which roughly captures the minimal number of edits necessary to transform the system output into an acceptable English translation of the source sentence. By comparing MT output with post-edited references, HTER provides more reliable estimates of translation quality than using translated references, especially at the segment level. The data for the analysis is drawn from an extended set of newswire reports in the 2008/2010 NIST Metrics for Machine Translation GALE Evaluation set[19]. For Chinese, there are 305 sentences (segments) translated to English by three different translation systems. For Arabic, there are 363 Arabic sentences (segments) translated by two systems.

The presence of discourse devices is analyzed only on the English side: the reference, the system hypothesis and its edited translation. Discourse connectives and their senses are identified using existing tools developed for English. Beyond its practical limitations, analyzing the reference interestingly reflects the choices made by the human translator: whether to choose to use a discourse connective, or to insert one to make an implicit relation on the source side explicit on the target side.

[19]Data used in this work includes more documents and the human edits not present in the official release.

Language	Sys1	Sys2	Sys3
ZH	0.097 (0.099)	0.117 (0.152)	0.144 (0.173)
AR	0.071 (0.148)	-0.089 (-0.029)	-

Table 5.1: Pearson (Spearman) correlation coefficient between lengths of source sentences and HTER values of three MT systems, for Chinese (ZH) and Arabic (AR). There is no strong relationship between sentence length and HTER values.

We first conduct analysis of variance (ANOVA) with HTER as dependent variable and the discourse factors as independent variables, and systems as subjects. We examine within-subject significance in each ANOVA model. For discourse factors that are significant at the 95% confidence level or higher according to the ANOVA analysis, we provide detailed breakdown of the system HTER for each value of the discourse factor.

In this paper we do not compare the performance of individual systems, but instead seek to understand if a discourse phenomena is problematic across systems.[20]

5.1.2 Sentence length and HTER

The presence of complex discourse structure is likely to be associated with longer sentences. It stands to reason that long sentences will be harder to process automatically and this reasoning has motivated the first approaches to text simplification (Chandrasekar et al., 1996). So before turning to the analysis of discourse phenomena, we examine the correlation between translation quality and sentence length. A strong correlation between the two would call for revival of interest in text simplification where syntactically complex sentences are transformed into several shorter sentences as a preprocessing step.

We find however that no strong relationship exists between the two, as shown by the correlation coefficients between HTER values and the number of words in each segment in Table 5.1.

Next we examine if sentence–discourse divergence between languages and the presence of certain discourse relations would be more indicative of the expected translation quality.

[20]For the readers with keen interest in system comparison, we note that according to ANOVA none of the differences in system performance on this data is statistically significant.

5.1.3 When a sentence becomes discourse

Some languages allow more information to be packed into a single sentence than is possible in another language, making single-sentence translations cumbersome and often ungrammatical. Chinese is known for sentences of this kind; for example, the usage of punctuation is very different in Chinese in the sense that a comma can sometimes function as a full stop in English, motivating a series of disambiguation tasks (Jin et al., 2004; Xue and Yang, 2011; Xu and Li, 2013). Special handling of long Chinese sentences were also shown to improve machine translation (Jin and Liu, 2010; Yin et al., 2007).

To investigate the prevalence of sentences in the source language (Chinese and Arabic in our case) that do not confirm to the notion of sentence in the target language (English for the purposes of this study), we separate the translation segments in the source language into two classes: a source sentence is considered 1-1 if the reference translation consists of exactly one sentence, and 1-many if the reference contains more than one sentence.

For Chinese, 26.2% of the source segments are 1-many. These sentences tend to be much longer than average (36.6% of all words in all reference translations are part of such segments). For Arabic, the numbers are 15.2% and 26.3%, respectively. Below is an example of a 1-many Chinese segment, along with the human reference and its translation by one of the systems:

> [**source**] 俄警方宣称，Erinys有一重要竞争对手RISC，利特维年科生前最后见面的人卢戈沃伊与友人都是从事这些行业。
>
> [**ref**] Russian police claim that Erinys has an important competitor RISC. The last people Litvinenko saw while he was alive, Lugovoi and his friends, were all engaged in these industries.
>
> [**sys**] Russian police have claimed that a major competitor, Litvinenko his last meeting with friends are engaged in these industries.

We conducted ANOVA on HTER, separately for each language, with type of segment (1-1 or 1-many) as the independent variable and systems treated as subjects. The test revealed that there is a significant difference in translation quality between 1-1 and 1-many segments for Chinese but not for Arabic. For the Chinese to English systems we further ran a Wilcoxon rank sum test to identify the statistical significance in performance for

AOV	Arabic	Chinese
$Pr(> F)$	0.209	0.0045*

	1-1	1-many
System	HTER	HTER
ZH-Sys1	16.22	19.03*
ZH-Sys2	19.54	21.02
ZH-Sys3	20.64	23.86*

Table 5.2: *Left:* ANOVA with type of segment (1-1 or 1-many) as independent variable and the three MT systems as subjects. *Right:* average HTER values for the three Chinese to English systems for 1-1 and 1-many segments. An asterisk ($*$) denotes significance at $p < 0.05$. 1-many segments is a significant factor in Chinese to English MT quality.

individual systems. For two of the three systems the difference is significant, as shown in Table 5.2.

We have now established that 1-many segments in Chinese to English translation are highly prevalent and their translations are of consistently lower quality compared to 1-1 segments. This finding suggests a cross language discourse analysis task of identifying Chinese sentences that cannot be translated into single English sentences. This task may be related to existing efforts in comma disambiguation in Chinese (Jin et al., 2004; Xue and Yang, 2011; Xu and Li, 2013) but the relationship between the two problems needs to be clarified in follow up work. Once 1-many segments are identified, source-side text simplification techniques may be developed (Siddharthan, 2006) to improve translation quality.

5.1.4 Explicit discourse relations

Explicit discourse relations such as COMPARISON, CONTINGENCY or TEMPORAL are signaled by an explicit connective, i.e., *however* or *because*. The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) provides annotations for the arguments and relation senses of one hundred pre-selected discourse connectives over the news portion of the Penn Treebank corpus (Marcus et al., 1993). Based on the PDTB, accurate systems for explicit discourse relation identification have been developed (Pitler and Nenkova, 2009; Lin et al., 2014). The accuracy of these systems is 94% or higher, close to human performance on the task. Here we study the influence of explicit discourse relations on machine translation quality and their interaction with 1-1 and 1-many segments.

AOV	Arabic	Chinese
$Pr(>F)$	0.39	0.0058*

(a) ANOVA with number of connectives (0, 1, more than one) as independent variable and the three MT systems as subjects.

% data (ZH)	No Conn	>1 Conn
all	53.77	15.08
1-many	13.77	5.25

(b) Proportion of reference Chinese sentences with no connective and more than one connective, for all segments and 1-many segments.

	all		1-many	
	No Conn	>1 Conn	No Conn	>1 Conn
ZH-Sys1	16.11	19.84^{+}	16.94	22.75^{+}
ZH-Sys2	19.96	22.39	20.47	23.25
ZH-Sys3	20.70	25.00*	22.30	29.68*

(c) Average HTER for the three Chinese-English systems, for reference translations with no connective and more than one connective, of the entire dataset and of 1-many segments.

Table 5.3: Number of discourse connectives and MT quality. An asterisk ($*$) or a plus ($+$) sign denotes significance at 95% and 90% confidence levels, respectively. Using more than one connective vs. no connective is a significant factor in Chinese to English MT quality.

5.1.5 Number of connectives

We identify discourse connectives and their senses (TEMPORAL, COMPARISON, CONTINGENCY or EXPANSION) in each reference segment using the system in Pitler and Nenkova (2009)[21]. We compare the translation quality obtained on segments with reference translation containing no discourse connective, exactly one discourse connective and more than one discourse connective.

The ANOVA indicates that the number of connectives is not a significant factor for Arabic translation, but significantly impacts Chinese translation quality. A closer inspection using Wilcoxon rank sum tests reveals that the difference in translation quality is statistically significant only between the groups of segments with no connective vs. those with more than one connective. Additionally, we ran Wilcoxon rank sum test over 1-1 and 1-many segments individually and find that the presence of discourse connectives is associated with worse quality only in the latter case. Effects above are illustrated in Table 5.3.

[21] http://www.cis.upenn.edu/∼epitler/discourse.html; We used the Stanford Parser (Klein and Manning, 2003).

AOV	Event	Arabic	Chinese
$Pr(>F)$	Contingency	0.61	0.028*
	Comp.:Temp.	0.047*	0.0041*

(a) ANOVA with relation sense as dependent variable and the three MT systems as subjects.

	Contingency	¬ Contingency	Comp.∧Temp.	¬(Comp.∧Temp.)
ZH-Sys1	20.15	16.72	23.58	16.64*
ZH-Sys2	21.69	19.80	26.16	19.63*
ZH-Sys3	25.87	21.16^{+}	27.20	21.21^{+}

(b) Average HTER for the three Chinese-English systems, for sentences containing a CONTINGENCY relation (6.89% of all data), without a CONTINGENCY relation, containing both COMPARISON and TEMPORAL relations (4.59% of all data) and without either of the two.

Table 5.4: Relation sense and MT quality. An asterisk (∗) or plus (+) sign denotes significance at 95% and 90% confidence levels, respectively. The presence of a CONTINGENCY relation is a significant factor in Chinese to English MT quality; the interaction between COMPARISON and TEMPORAL is significant for both Chinese and Arabic.

5.1.6 Relation senses

Here we study whether discourse relations of specific senses pose more difficulties on translations than others and whether there are interactions between senses. In the ANOVA analysis we used a binary factor for each of the four possible senses. For example, we compare the translation quality of segments that contain COMPARISON relations in the reference translation with those that do not.

The relation sense makes a significant difference in translation quality for Chinese but not for Arabic. For Chinese specifically sentences that express CONTINGENCY relations have worse quality translations than sentences that do not express CONTINGENCY. One explanation for this tendency may be that CONTINGENCY in Chinese contains more ambiguity with other relations such as TEMPORAL, as tense is expressed lexically in Chinese (no morphological tense marking on verbs). Finally, the interaction between COMPARISON and TEMPORAL is significant for both languages.

Table 5.4 shows the effect of relation sense on HTER values for Chinese.

5.1.7 Human edits of discourse connectives

A relation expressed implicitly without a connective in one language may need to be explicit in another. Moreover, the expressions themselves are used differently; for example, the

paired connective "虽然...但是" (despite...but) in Chinese should not be translated into two redundant connectives in English. It is also possible that the source language contains an explicit discourse connective which is not translated in the target language, as has been quantitatively studied recently by Meyer and Webber (2013). An example from our dataset is shown below:

> [**source**] 还有些人可到大学的游戏专业深造，而后被聘请为大游戏厂商的技术顾问等。
>
> [**ref**] Still some others can receive further professional game training in universities and later(*Temporal*) be employed as technical consultants by large game manufacturers, etc.
>
> [**sys**] Some people may go to the university games professional education, which is appointed as the big game manufacturers such as technical advisers.
>
> [**edited**] Some people may go to university to receive professional game education, and later(*Temporal*) be appointed by the big game manufacturers as technical advisers.

The system fails to translate the discourse connective "而后" (later), leading to a probable misinterpretation between receiving education and being appointed as technical advisors.

Due to the lack of reliable tools and resources, we approximate mismatches between discourse expressions in the source and MT output using discourse-related edits. We identify explicit discourse connectives and their senses in the system translation and the human edited version of that translation. Then we consider the following mutually exclusive possibilities: *(i)* there are no discourse connectives in either the system output or the edit; *(ii)* the system output and its edited version contain exactly the same discourse connectives with the same senses; *(iii)* there is a discourse connective present in the system output but not in the edit or vice versa. In the ANOVA we use a factor with three levels corresponding to the three cases described above. The factor is significant for both Chinese and Arabic. In both languages, the mismatch case *(iii)* involves significantly higher HTER than either case *(i)* or *(ii)*. The human edit rate in the mismatch class is on average four points greater than that in the other classes.

Obviously, the mismatch in implicit/explicit expression of discourse relation is related to the first problem we studied, i.e., if the source segment is translated into one or multiple sentences in English, since discourse relations between adjacent sentences are more often

	% data			AOV
	Mismatch	Mismatch (1-1)	¬Mismatch (1-1)	$Pr(>F)$
Arabic	21.27	15.47	69.34	4.0×10^{-6}*
Chinese	29.51	17.05	56.82	4.1×10^{-11}*

(a) Columns 2-4: percentage of segments with mismatches, 1-1 segments with mismatches, and 1-1 segments without mismatches. Column 5: ANOVA with mismatch type as independent variable and the three MT systems as subjects.

	¬Mismatch	Mismatch	¬Mismatch (1-1)	Mismatch (1-1)
AR-Sys1	11.23	15.92*	10.86	16.24*
AR-Sys2	11.64	15.74*	11.58	16.65*
ZH-Sys1	15.57	20.72*	15.47	19.13*
ZH-Sys2	19.02	22.34*	18.68	22.52*
ZH-Sys3	11.64	15.74*	19.57	26.07*

(b) Average HTER for Chinese (ZH) and Arabic (AR) segments where there is no mismatch vs. there is a mismatch, for all segments and 1-1 segments only.

Table 5.5: The impact of discourse connective mismatch between human edits and system translations on MT quality, for 1-1 and 1-many segments. An asterisk (*) denotes significance at $p < 0.05$. This mismatch is a significant factor in MT quality from both Chinese and Arabic to English.

implicit (than intra-sentence ones). For this reason we performed a Wilcoxon rank sum test for the translation quality of segments with discourse mismatch conditioned on whether the segment was 1-1 or 1-many. For both languages a significant difference was found for 1-1 sentences but not 1-many. Table 5.5 shows the proportion of data in each of the conditioned classes and the average HTER for sentences from the mismatch case *(iii)* where a discourse connective was edited and the others (no such edits). Translation quality degrades significantly for all systems for the mismatch case, over all data as well as 1-1 segments.

5.1.8 Discussion and conclusion

We showed that translation from Chinese to English is made more difficult by various discourse events such as the use of discourse connectives and the type of relations they signal. None of these discourse factors has a significant impact on translation quality from Arabic to English. Translation quality from both languages is adversely affected by translations of discourse relations expressed implicitly in one language but explicitly in the other or by paired connectives. Our experiments indicate that discourse usage may affect machine

translation between some language pairs but not others, and for particular relations such as CONTINGENCY. Finally, we established the need to identify sentences in the source language that would be translated into multiple sentences in English. Especially in translating from Chinese to English, there is a large number of such sentences which are currently translated much worse than other sentences. In the next section, we will focus on the identification of these sentences in Chinese. For Arabic, these sentences are not linked with significantly worse machine translation quality, hence we will not further the discussion on Arabic. The very different results presented here regarding Chinese and Arabic opens future directions to explore the organization of sentences in the two languages, and whether it is related to previously discovered syntactic differences in the context of machine translation (Marton and Resnik, 2008).

5.2 Discourse vs. sentence: identifying content-heavy sentences

To generate text, people and machines need to decide how to package the content they wish to express into clauses and sentences. There are multiple possible renderings of the same information, with varying degrees of ease of comprehension, compactness and naturalness. Some sentences, even though they are grammatical, would be more accessible to a reader if expressed in multiple sentences. We call such sentences *content heavy* sentences, or *heavy sentences* for brevity.

In the established areas of language research, text simplification and sentence planning in dialog and generation systems are clearly tasks in which identification of content-heavy sentences is of great importance. In this paper we introduce a novel flavor of the task in the cross-lingual setting, which in the long term may guide improvements in machine translation. We seek to identify sentences in Chinese that would result in heavy sentences in English if translated to a single sentence.

Example I in Table 5.6 shows a Chinese sentence and its two English translations A and B. Translator A used three English sentences to express all the information. Translator B, on the other hand, used a single sentence, which most readers would find more difficult to

[Example I] 虽然菲军方在南部的巴西兰岛上部署了5000多兵力，并在美军的帮助下围剿阿布沙耶夫分子，但迄今收效不大。*Although the Philippine army on the southern Basilan island deployed over 5,000 troops, and with the US army's help are hunting down ASG members, but so far achieved little.* **[A]** The Philippine army has already deployed over 5 thousand soldiers on the southern island of Basilan. With the help of U.S. army, these soldiers are searching and suppressing members of Abu Sayyaf. However, there is not much achievement this far. **[B]** The Philippine military has stationed over 5,000 troops on Basilan Island in the southern Philippines and also tried to hunt down ASG members with the help of the United States, yet so far it has little success.	**[Example II]** 端粒是染色体末端的结构，随着细胞老化和失去分裂能力，端粒会逐渐缩短长度，换言之，端粒愈长显示细胞老化愈慢。*Telomeres are chromosome ends structures, with cell aging and losing division ability, telomeres will gradually decrease length, in other words, telomeres the longer shows cell aging the slower.* **[A]** Telomeres are structures at the ends of chromosomes, which gradually reduce in length with the aging of the cells and their loss of the ability to divide. In other words, longer telomeres indicate the slower aging of the cells. **[B]** Telomeres are the physical ends of chromosomes. As cells age and lose the ability to divide, the telomeres shrink gradually. That is to say, longer telomeres indicate that cells are aging more slowly.

Table 5.6: Examples of Chinese sentences expressed in multiple English sentences.

read. Example II illustrates a case where a translator would be hard pressed to convey all the content in a sentence in Chinese into a single grammatical English sentence.

Here we provide an operational characterization of content-heavy sentences in the context of Chinese-English translation. Instead of establishing guidelines for standalone annotation, we repurpose datasets developed for evaluation of machine translation consisting of multiple reference translations for each Chinese sentence. In this cross-lingual analysis sentences in Chinese are considered content-heavy if their content would be more felicitously expressed in multiple sentences in English.

We first show that with respect to English, content-heavy Chinese sentences are common. A fifth to a quarter of the sentences in the Chinese news data that we analyze are translated to multiple sentences in English. Moreover our experiments with reader preference indicate that for these sentences, readers strongly prefer multi-sentence translation to a single-sentence translation. We also compare the difference in machine translation quality for heavy sentences and find that it is considerably lower than overall system performance.

We study the connection between heavy sentences and the factors used in prior work to split a Chinese sentence into multiple sentences, showing that they do not fully determine the empirically defined content-heavy status. Furthermore we present an effective system to automatically identify content-heavy sentences in Chinese.

5.2.1 Data

In this work we use three news datasets: the newswire portion of the NIST 2012 Open Machine Translation Evaluation (OpenMT) (Group, 2013), Multiple-Translation Chinese (MTC) parts 1-4 (Huang et al., 2002; Huang et al., 2003; Ma, 2004; Ma, 2006), and the Chinese Treebank (Xue et al., 2005). In OpenMT and MTC, multiple reference translations in English are available for each Chinese segment (sentence).

To study the relationship between content-heavy sentences and reader preference for multi-sentence translations (Section 5.2.2), we use OpenMT (688 segments) and MTC parts 2-4 (2,439 segments), both of which provide four English translations for each Chinese segment. This analysis forms the basis for labeling heavy sentences for supervised training and evaluation (Sections 5.2.4, 5.2.5, 5.2.6).

The Chinese Treebank (CTB) has been used in prior work as data for identifying full-stop commas. Moreover, 52 documents in MTC part 1 were drawn from the CTB. The intersection of the two datasets allows us to directly analyze the relationship between heavy sentences and full-stop commas in Chinese (Section 5.2.4). Furthermore we use this intersection as test set to identify heavy sentences so we can directly compare with models developed for comma disambiguation. To be consistent with the rest of the MTC data, we use 4 out of the 11 translators in part 1 in these experiments.[22]

Our model for Chinese full-stop comma recognition is trained following the features and training sets specified in Xue and Yang (2011)[23], excluding the overlapping MTC/CTB documents mentioned above. There are 12,291 sentences in training that contain at least one comma. A classifier for detecting heavy sentences is trained on OpenMT and MTC (excluding the test set). A quick inspection of both datasets reveals that Chinese sentences without a comma were never translated into multiple sentences by more than one translator. Therefore in our experiments we consider only sentences that contain at least one comma. There are 301 testing sentences, 511 training sentences in OpenMT and 2418 in MTC. Sentences are processed by the Stanford NLP packages[24]. CTB gold-standard parses are

[22] We did not use translator IDs as parameters in any of our systems.

[23] Document IDs 41-325, 400-454, 500-554, 590-596, 600-885, 900, 1001-1078, 1100-1151.

[24] The Stanford segmenter (Tseng et al., 2005), parser (Levy and Manning, 2003) and the CoreNLP package (Manning et al., 2014)

	OpenMT		MTC	
# ref multi	% data	% best multi	% data	% best multi
0	65.4	0	58.9	0
1	7.4	23.5	20.4	20.1
2	7.0	66.7	8.3	56.7
3	9.2	88.9	7.9	89.6
4	11.0	100	4.6	100

Table 5.7: Percentage of Chinese sentences for which a given number of translators (# ref multi) prefer to use multiple sentences in English (% data), along with percentage of times a multi-sentence translation was selected as most fluent and comprehensible by readers (% best multi).

used to obtain full-stop commas and to train comma disambiguation models.

5.2.2 Content-heavy sentences: definition

First we quantify how often translators choose to translate a Chinese sentence into multiple English sentences. Content-heavy Chinese sentences are those for which there is a strong preference to produce multiple sentences when translating to English (at the end of the section we present specific criteria).

Obviously, splitting a sentence into multiple ones is often possible but is not necessarily preferred. In Table 5.7, we show in the "%data" columns the percentage of source sentences split in translation by 0, 1, 2, 3 and all 4 translators. For about 20% of segments in OpenMT and 15% in MTC, at least three of the translators produce a multi-sentence translation, a rate high enough to warrant closer inspection of the problem.

Next, we conduct a study to find out what level of translator agreement leads to strong reader preference for the same information to be presented in multiple sentences.

For each Chinese segment with one, two or three multi-sentence reference translations, we ask five annotators on Mechanical Turk to rank the reference translations according to their general flow and understandability. The annotators saw only the four randomly ordered English translations and were not shown the Chinese original, with the following instruction:

> Below are 1-2 sentence snippets that describe the same content. Some are more readable and easier to understand than others. Your task is to rank them from the best to worst

in terms of wording or flow (organization). There can be ties, but you have to pick one that is the best.

We obtain reader preference for each segment in the following manner: for each annotator, we take the highest ranked translation and check whether it consists of multiple sentences. In this way we have five binary indicators. We say readers prefer a sentence to have a multi-sentence translation in terms of flow and comprehensibility if the majority of these five indicators are positive.

In the "%best multi" columns of Table 5.7, we tabulate the percentage of segments with majority preference for multi-sentence translation, stratified by the number of translators who split the content. Obviously the more multi-sentence translations there are, the higher the probability that the readers will select one as the best translation. We are interested in knowing for which conditions the preference for multi-sentence translation exceeds the probability of randomly picking one.

When only one (out of four) translations is multi-sentence, the best translations chosen by the majority of readers contain multiple sentences less often than in random selection from the available translations. When two out of the four reference translations are multi-sentence, the reader preference towards them beats chance by a good margin. The difference between chance selection and reader preference for multiple sentences grows steadily with the number of reference translations that split the content. These data suggest that when at least two translators perform a multi-sentence translation, breaking down information in the source sentence impacts the quality of the translation.

Hence we define content-heavy sentences in Chinese to be those for which at least two out of four reference translations consist of multiple sentences.

5.2.3 A challenge for MT

We now quantitatively show that heavy sentences are particularly problematic for machine translation. We collect translations for each segment in OpenMT and MTC from the Bing Translator. We split the sentences into two groups, heavy and other, according to the gold standard label explained in the previous section. We then compare the BLEU score for sentences in a respective group, where each group is in turn used as a test set. The difference

Criteria	%data(Y)	bleu(Y)	bleu(N)	Δbleu
heavy	27.2	15.34	19.24	**3.9**

Table 5.8: Percentage of content-heavy Chinese sentences, along with BLEU scores for heavy and non-heavy sentences and their difference. The BLEU score for content-heavy sentences are much lower.

heavy	fs-comma	No fs-comma
N	19	180
Y	40	62

Table 5.9: Counts of heavy (Y) and non-heavy (N) sentences with and without full-stop commas.

in BLEU scores (Δbleu) is a strong indicator whether these sentences are challenging for MT systems.

In Table 5.8 we show the BLEU scores and Δbleu for sentences that are heavy (Y) and non-heavy (N). Also included in the table is the percentage of heavy sentences in all the data.

Translations for heavy sentences received a BLEU score that is 3.9 points lower than those that are not. This clearly illustrates the challenge and potential for improvement for MT systems posed by content-heavy sentences. Therefore the ability to reliably recognize them provides a first step towards developing a better translation approach for such sentences.

5.2.4 Comma usage and heavy sentences

In Chinese, commas can sometimes act as sentence boundaries, similar to the function of an English period. In Xue and Yang (2011), the authors showed that these full-stop commas can be identified in the constituent parse tree as coordinating IPs at the root level, shown in Figure 5.1. Fancellu and Webber (2014) demonstrated that it is beneficial to split sentences containing negation on these types of commas, translate the resulting shorter sentences separately, then stitch the resulting translations together. They report that this approach prevented movement of negation particles beyond their scope. Here we study the degree to which the content-heavy status of a sentence is explained by the presence of a full-stop comma in the sentence. We show that they are interrelated but not equivalent.

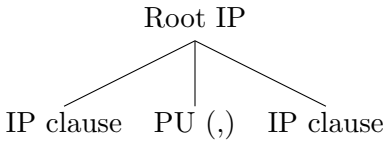


Figure 5.1: Commas separating coordinating IPs at the root. These full-stop commas can act as sentence boundaries in Chinese (Xue and Yang, 2011).

Corpus analysis. First we study how often a heavy sentence contains a full-stop comma and vice versa, using the overlapping MTC/CTB documents. We show in Table 5.9 the number of heavy and non-heavy sentences with and without full-stop commas[25]. When there is a full-stop comma in the sentence, there is a higher chance that the sentence is content-heavy. Yet of the 102 heavy sentences in this data, fewer than 40% contain full-stop commas; of the 242 sentences without full-stop commas, more than a quarter are heavy. Therefore, although comma usage in the Chinese sentence may provide clues for detecting content heaviness, the two phenomena are not equivalent and heavy sentences are not fully explained by the presence of full-stop commas.

Learning with full-stop commas. Here we evaluate the usefulness of using full-stop commas as training data to predict whether a sentence is content-heavy. From the analysis presented above we know that the two tasks are not equivalent. Nevertheless we would like to test directly if the Chinese Treebank—the large (but noisy for the task at hand) data available for comma function disambiguation—would lead to better results than learning on the cleaner but much smaller datasets for which multiple translations are available.

We use logistic regression as our classification model[26]. The performance of identifying heavy sentences on the MTC/CTB overlapping test set is compared using the following methods:

- **[Parallel]**: A classifier is trained using four English translations for each Chinese sentence (OpenMT and MTC training set). Following the definition in Section 5.2.2, content-heavy sentences are those translated into multiple English sentences by two or more translators.

[25]For the study we exclude sentences without a comma. A χ^2 test for the strength of association between the presence of full stop commas and heavy sentence status shows high significance.

[26]We use the Liblinear package Fan et al. (2008).

Training	Accuracy	Precision	Recall
parallel	**75.75**	**69.86**	**50**
oracle comma	73.09	67.8	39.2
predicted comma	74.42	66.67	49.02

Table 5.10: Performance to identify heavy sentences using multiple reference data (parallel) vs. full-stop comma oracle labels (oracle comma) and predicted full-stop commas (predicted comma). It is more advantageous to learn from multiple reference data.

- [**Oracle comma**]: A test sentence is assigned to class "heavy" if there is a full-stop comma in its corresponding gold standard parse tree.
- [**Predicted comma**]: We train a comma disambiguation system on CTB to predict if a comma is a full-stop comma. In testing, a sentence is marked "heavy" if it contains a predicted full-stop comma.

Features. We reimplemented the per-comma features used in Xue and Yang (2011)[27]. As in their best performing system, features are extracted from gold-standard parse trees during training and from automatic parsing during testing. These include: words and part-of-speech tags immediately before and after the comma; left- and right-sibling node labels of the parent of the comma; ordered ancestor node labels above the comma; punctuation tokens ordered from left to right of the sentence; whether the comma has a coordinating IP structure; whether the comma's parent is a child of the root of the tree; whether there is a subordination before the comma; whether the difference in number of words before and after the comma is greater than or equal to seven.

For *parallel*, feature values are accumulated from all the commas in the sentence. For binary features, we use an *or* operation on the feature values for each individual comma.

Results and comparison. In Table 5.10, we show the accuracy, precision and recall for identifying content-heavy sentences using the three methods described above. We do not include the majority baseline here because it assumes no sentences are content heavy.

Interestingly, the system using oracle information in each test sentence for full-stop commas performs the worst. The system trained to identify full-stop commas outperform the

[27]For *predicted comma*, our reimplementation of Xue and Yang (2011) gave practically identical results to those reported in the original paper on the test set that they used.

oracle system with about 10% better in recall and less than 1% lower in precision. This finding strongly suggests that the features used for learning capture certain characteristics of heavy sentences even with non-ideal training labels. The best performance is obtained learning directly on parallel corpora with multiple reference translations. Note that we try to provide the best possible setting for full-stop comma prediction, using much more training data, gold-standard parses, same-domain training and testing, as well as the reimplementation of state-of-the-art system. These settings allow us to conservatively interpret the results listed here, which confirm that content-heaviness is different from using a full-stop comma in the Chinese sentence. It is more advantageous—leading to higher precision and overall accuracy—to learn from data where translators encode their interpretation in the form of multi-sentence translations.

5.2.5 Features to characterize content-heavy sentences

In this section, we experiment with a wide range of features from the sentence string, part-of-speech tags and dependency parse trees.

Baseline. Intuitively, sentence length can be an indication of too much content that needs to be repackaged into multiple sentences. Therefore as our baseline we train a decision tree using the number of words in a Chinese sentence.

Sentence structure cues. We collect potential signals for structural complexity: punctuation, conjunctions, prepositional phrases and relative clauses. As features we count the number of commas, conjunction, preposition and postposition part-of-speech tags. In Chinese "DE" often marks prepositional phrases or relative clauses among other functions (Chang et al., 2009a). Here we include a simple count the number of "DEG" tags in the sentence.

Dependencies. Dependency grammar captures both syntactic and semantic relationship between words and are shown to improve reordering in MT (Chang et al., 2009b). To account for such relational information we include two feature classes: the percentage of each dependency type and the typed dependency pairs themselves. For the latter we use

the universal part-of-speech tags (Petrov et al., 2012) for each word rather than the word itself to avoid too detailed and sparse representations. For example, the relation *dobj(处理/handle, 事情/matter)* becomes feature *dobj(verb, noun)*.

Furthermore, we use dependency trees to extract four features for potentially complex constructions. First, we indicate the presence of noun phrases with heavy modifiers on the left. These are frequently used in Chinese and would require a relative clause or an additional sentence in English. Specifically we record the maximum number of dependents for the nouns in the sentence. The second type of construction is the use of serial verb phrases, such as $VP \rightarrow VP\ PU\ VP$. We record the number of dependents of the head verb of the sentence. The third feature class is the typed dependencies (over universal POS tags) whose edge crosses a comma. Finally, we also record the maximum number of dependents in the sentence to capture the general phrasal complexity in the sentence.

Parts-of-speech. POS information captures numerous aspects of the sentence such as the frequency of different classes of words used and the transition between them. Historically they are also shown to be helpful for phrase boundary detection (Taylor and Black, 1998). Here, we first convert all Chinese POS tags into their corresponding universal tags. We then use the percentage of each tag and tag bigram as two feature classes. To capture the transition of each phrase and clause in the sentence, we construct functional POS trigrams for each sentence by removing all nouns, verbs, adjectives, adverbs, numbers and pronouns in the sentence. Percentages of these sequences are used as feature values.

Comma disambiguation features. We also incorporate most of the features proposed by Xue and Yang (2011), aggregated in the same way as the *parallel* method (cf. Section 5.2.4). These include: POS tags immediately before and after the comma; left- and right-sibling node labels of the parent of the comma; the punctuation tokens ordered from left to right in the sentence, whether the comma has a coordinating IP structure; whether the comma's parent is a child of the root of the tree; whether there is a subordination before the comma; whether the difference in number of words before and after the comma is greater than or equal to seven.

Features	Training	Accuracy	Precision	Recall
baseline	MTC+OpenMT	71.43	73.5	24.5
full set	OpenMT	76.41	66.67	**60.78**
full set	MTC	78.41	74.03	55.9
full set	MTC+OpenMT	**80.73**	**79.73**	57.84

Table 5.11: Accuracy, precision and recall of classifying content-heavy sentences, using MTC and/or OpenMT as training data. *Baseline*: sentence length; *full set*: full set of features proposed in our work.

5.2.6 Recognizing content-heavy sentences

We train a logistic regression model as in the *parallel* method in Section 5.2.4 using features illustrated above. In Table 5.11, we show the performance of detecting heavy sentences using four systems: the baseline system using the number of words in the sentence and three systems using our full feature set, trained on MTC, OpenMT and both.

The baseline performance is characterized by a remarkably poor recall. It becomes apparent that length alone cannot characterize content-heaviness. On the other hand, using the full feature set achieves an accuracy of above 80%, a precision close to 80% and a recall about 58%. The improvement in precision and recall over using oracle full-stop commas (Table 5.10) are about 12% and 19%. When compared with using features tuned for comma disambiguation from Xue and Yang (2011) (Table 5.10), our full feature set achieved a 5% increase in accuracy, about 10% increase in precision and 8% increase in recall.

We also demonstrate the usefulness of having more multi-reference translation data by comparing training using MTC and OpenMT individually and both. Remarkably, using only the very small dataset of OpenMT is sufficient to produce a predictor that is more accurate than all of the methods listed in Section 5.2.4. Adding these examples to MTC drastically improves precision by more than 13% with a less than 3% drop on recall.

Finally, we consider the portions of our test set for which at least n translators provided a multi-sentence translation (n ranges from 0 to 4). In Table 5.12 we show the respective precision, recall and the average posterior probability from the classifier for marking a sentence as content-heavy. The recall values are in general lower due to a skewed class distribution (the minority class is content-heavy). There is a clear trend that the classifier is more confident and has higher precision for sentences where more translators produce multi-

#ref multi	≥ 0	≥ 1	≥ 2	≥ 3	4
#seg	301	187	102	58	25
precision	79.73	84.29	100	100	100
recall	57.85	57.84	57.84	68.98	76
posterior	0.29	0.40	0.53	0.61	0.67

Table 5.12: Number of segments, precision, recall and posterior probability for examples where at least 0, 1, 2, 3 or 4 translators split the sentence. When more translators split the sentence, the classifier is more confident and achieves better performance.

sentence translations. Although the model is not highly confident in all groups, the precision of the predictions are remarkably high. Miss rate also decreases when more translators translate the source into multiple sentences.

Post-hoc feature analysis Here we identify which of the feature classes from our full set are most helpful by performing forward feature selection: in each iteration, the feature class that improves accuracy the most is selected. The process is repeated until none of the remaining feature classes leads to improvement when added to the model evaluated at the previous iteration. We use our test data as the evaluation set for forward selection, but we do so only to evaluate features, not to modify our system.

Five feature classes are selected using this greedy procedure. The first selected class is the typed dependencies over universal POS tags. Remarkably, this single feature class achieves 76.6% accuracy, a number already reasonably high and better than features used in Xue and Yang (2011). The second feature added is whether there is a comma of coordinating IP structure in the automatic parse tree of the sentence. It gives a further 1.7% increase in accuracy, showing that the comma structure provide useful information as features for detecting heavy sentences. Note that this feature does not represent full stop commas, i.e., it does not record whether the comma is under the root level of the parse tree. The next selected class is typed dependencies over universal POS tags that have an edge across commas in the sentence, with an 1% increase in accuracy. The fourth feature selected is the number of prepositions and postposition POS tags in the sentence, improving the accuracy about 1%. Finally, part-of-speech tags before each comma are added, with a 0.3% improvement of accuracy.

The results from forward selection analysis reveal that the dependency structure of

a sentence captures the most helpful information for heavy sentence identification. The interplay between punctuation and phrase structure gives further important enhancements to the model. The final accuracy, precision and recall after forward selection are 0.804, 0.8209, 0.5392, respectively. This overall performance shows that forward selection yields a sub-optimal feature set, suggesting that the other features are also informative.

5.2.7 A challenge for MT: revisited

It is important to know whether a predictor for content-heavy sentences is good at identifying challenging sentences for applications such as machine translation. Here, we would like to revisit Section 5.2.3 and see if *predicted* heavy sentences are harder to translate.

For all the source sentences in OpenMT and MTC, we compare five criteria for dividing the test data in two subsets: whether the sentence contains a full-stop comma or not; whether the sentence is longer than the baseline decision tree threshold (47 words) or not; whether the sentence is predicted to be content-heavy with posterior probability threshold of 0.5, 0.55 and 0.6. Predictions for the training portion is obtained using 10-fold cross-validation. In the same manner as Table 5.8, Table 5.13 shows the percentage of data that satisfies each criterion, BLEU scores of Bing translations for sentences that satisfy a criterion and those that do not, as well as the difference of BLEU between the two subsets (Δbleu). As reference we also include numbers listed in Table 5.8 using oracle content-heavy labels.

First, notice that regardless of the posterior probability threshold, the numbers of sentences predicted to be content-heavy are much larger than that using the length cutoff. These sentences are also collectively translated much worse than the sentences in the other subset. Sentences that contain a predicted full-stop comma are also harder to translate, but show smaller difference in BLEU than when sentence heaviness or length are used as separation criterion. As the posterior probability threshold goes up and the classifier becomes more confident when it identifies heavy sentences, there is a clear trend that system translations for these sentences become worse. These BLEU score comparisons indicate that our proposed model identifies sentences that pose a challenge for MT systems.

Criteria	%data(Y)	bleu(Y)	bleu(N)	Δbleu
fs-comma	21.6	16.01	18.43	2.42
length threshold	8.6	15.38	18.3	2.92
pred-heavy (0.5)	22.72	15.81	18.77	2.96
pred-heavy (0.55)	19.72	15.47	18.76	3.29
pred-heavy (0.6)	16.67	14.95	18.77	3.82
oracle heavy	27.4	15.34	19.24	3.9

Table 5.13: For each criterion to separate heavy and non-heavy sentences, the percentage of heavy sentences (*%data(Y)*), BLEU scores for heavy and non-heavy sentences, and their differences. The criteria are: *fs-comma*: whether the sentence contains a full-stop comma; *length threshold*: whether it is longer than the length threshold; *pred-heavy (prob)*: whether it is predicted predicted content heavy with the posterior probability cutoff *prob*; *oracle heavy*: whether it is content heavy according to the oracle definition. Our system can more reliably identify sentences that are harder to translate.

5.2.8 Conclusion

In this work, we propose a cross-lingual task of detecting content-heavy sentences in Chinese, which are best translated into multiple sentences in English. We show that for such sentences, a multi-sentence translation is preferred by readers in terms of flow and understandability. Content-heavy sentences defined in this manner present practical challenges for MT systems. We further demonstrate that these sentences are not fully explained by sentence length or syntactically defined full-stop commas in Chinese. We propose a classification model using a rich set of features that effectively identify these sentences.

5.3 Discussion and future work

Differences in the distribution of content into sentences cataloged in this chapter point out a definite issue in different languages currently under-investigated in text-to-text generation systems. One possible way to improve MT systems is to incorporate sentence simplification before translation (Mishra et al., 2014). Future work could use our proposed model to detect heavy sentences that need such pre-processing. Our findings can also inspire informative features for sentence quality estimation, in which the task is to predict the sentence-level fluency (Beck et al., 2014). We have shown that heavy Chinese sentences are likely to lead to hard to read, cumbersome sentences in English.

Another important future direction lies in text simplification. In our inspection of

parallel Wikipedia/Simple Wikipedia data (Kauchak, 2013), around 23.6% of the aligned sentences involve a single sentence on one side and multiple sentences on another. Naturally, not all sentences in a text need to be simplified; for example, the Simple Wikipedia preserved many sentences from the original Wikipedia. In Section 7.1, we discuss text specificity in identifying sentences that need simplification in the first place. Ideas from this work can also be useful to this task.

Chapter 6

The organization of specific information

So far we have discussed the organization of general and specific information in two aspects: the INSTANTIATION discourse relation, and the specificity of textual units such as sentences and words. Separately, we have examined content-heavy sentences—sentences in Chinese that need to be translated into multiple English sentences. In this chapter, we investigate the connection between specificity, INSTANTIATION and content-heavy sentences.

We found that per-word specificity in content-heavy sentences is higher than non-heavy sentences. Meanwhile, multi-sentence translations of content-heavy sentences are overall less specific than single sentence translations. Later in Section 7.1, we will show that high specificity is closely related to sentences that need simplification, so doing a multi-sentence translation can be viewed as a way of enhancing the intelligibility of the translated text. Furthermore, specificity complements our classifier based on syntactic patterns (Section 5.2) to predict whether a sentence is content-heavy.

For INSTANTIATION, we seek adjacent Chinese sentence pairs whose translations have similar characteristics as the implicit INSTANTIATION relation in English. To do this, we use the classifier presented in Section 3.2 on the reference translations between adjacent Chinese sentences. Among the predicted second arguments of INSTANTIATION, more than half of them are content-heavy sentences, showing a strong association between the two.

In an analysis of how a sentence is split into a multi-sentence translation, it is clear that

a content-heavy sentence is most popularly split into equally general/specific segments, less often into specific-general segments, and much less often general-specific segments. Moreover, explicit discourse relation profiles are different when the arguments are between whole Chinese sentences, between split components and within a split component. This suggests that the translator is guided to some extent by explicit discourse relations when doing a multi-sentence translation. In addition, it is not likely that a translator split a sentence such that the latter part elaborates the first.

6.1 Data and settings

Data. We use the same multiple translation data from Section 5.2: Multiple Translation Chinese (MTC) and OpenMT. Each Chinese sentence has 4 reference translations. There are 3.5K Chinese sentences in 487 documents. As in Section 5.2, A Chinese sentence is heavy if at least two translators did a multi-sentence translation.

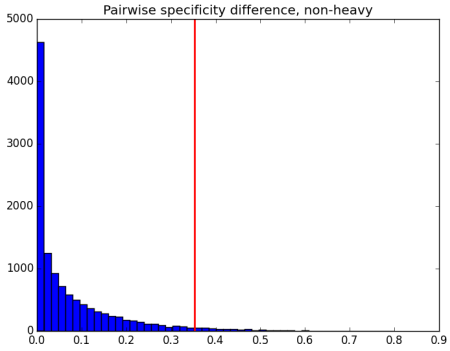
Specificity. We measure the specificity of the reference translations predicted using SPECITELLER (Section 3.3). If there are multiple sentences in the translation of one Chinese sentence, we weight the final score by the number of words in each sentence as in Li and Nenkova (2015b) (Section 7.1):

$$spec(\text{ref}) = \frac{1}{\sum_{s \in \text{ref}} |s|} \sum_{s \in \text{ref}} |s| \times \text{Speciteller(s)} \tag{6.1}$$

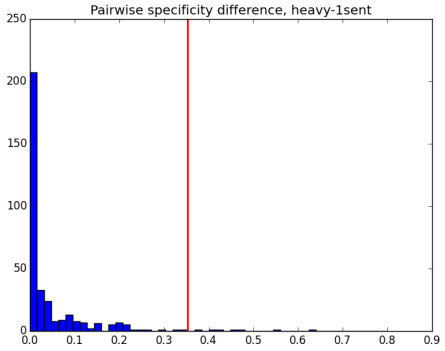
where s denotes a sentence in the reference translation *ref*, $|s|$ denotes the length of the sentence.

A note on sentence length. The relatively high correlation between sentence length and specificity (discussed in Section 3.3.6) means that we need to account for sentence length for the analysis we conduct here. To this end, we take two measures: specificity bucketed by sentences with similar lengths, and specificity normalized by the number of words in the sentence. If there are multiple sentences, we use the average.

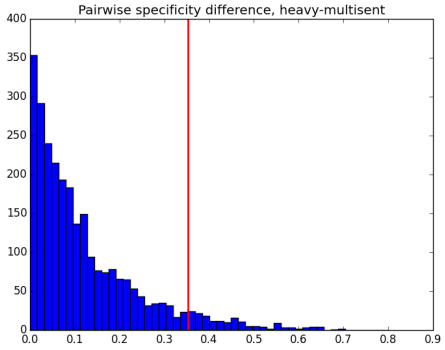
Discourse relations. For explicit discourse relations, we run the NUS discourse parser (Lin et al., 2009) on each article's four reference translations. For the implicit INSTANTIA-



(a) One sentence translations of non-heavy sentences.



(b) One sentence translations of content-heavy sentences.



(c) Multi-sentence translation of content-heavy sentences.

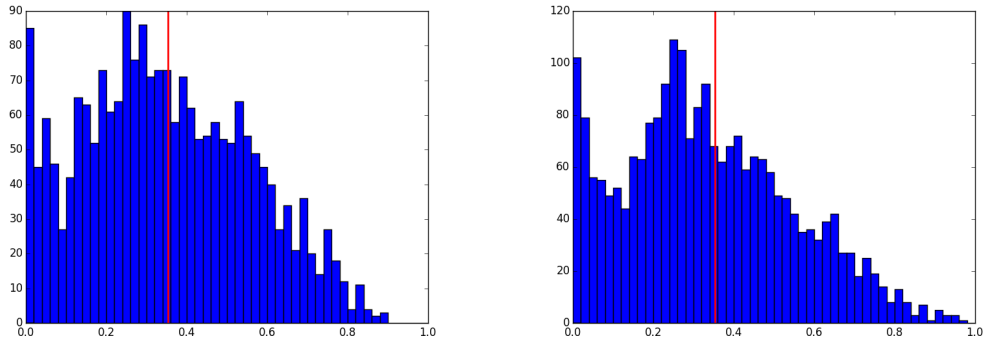
Figure 6.1: Histograms of pairwise specificity differences of reference translations of the same Chinese sentence; showing only pairs where both translations are single-sentence or multi-sentence. X-axis: pairwise specificity difference; y-axis: number of sentence pairs. Red lines indicate the average difference between 1000 randomly selected sentences. For the same source sentence, the specificity of translations of the same type (either one- or multi-sentence) is much more consistent compared to those of different types (Figure 6.2).

TION relation, we use our model in Section 3.2 on each article's four reference translations.

6.2 Specificity

6.2.1 Consistency across reference translations

First we study the distribution of, and differences in, specificity of the translations of content-heavy sentences. To do this we inspect the consistency of specificity scores between reference translations of the same Chinese sentence.



(a) Non-heavy sentences.

(b) Heavy sentences.

Figure 6.2: Histograms of pairwise specificity differences of reference translations of the same Chinese sentence; showing only pairs where one translation is single-sent and the other is multi-sent. X-axis: pairwise specificity difference; y-axis: number of sentence pairs. Red lines indicate the difference between 1000 randomly selected sentences. For the same source sentence, the specificity of translations of different types (one- or multi-sentence) is much more different compared to those of the same type (Figure 6.1).

We plot the pairwise specificity difference of the reference translations and consider two situations. Let r_i, r_j be two translations of a Chinese sentence s. Figure 6.1 shows pairwise specificity differences when both r_i and r_j are one-sentence translations or both of them are multi-sentence translations. In Figure 6.2 we plot pairwise specificity differences when either r_i or r_j is a single-sentence translation while the other is a multi-sentence translation. For reference, the average pairwise specificity differences between 1,000 randomly selected reference sentences is 0.353 ($\sigma = 0.293$), which is shown as a red line in each figure.

The specificity of one-sentence translations of the same source sentence (Figure 6.1) is highly consistent, regardless of whether the source sentence is content-heavy or not. There is slightly more variation among multi-sentence translations of heavy sentences, but they are still overwhelmingly below that of two random sentences. On the other hand, variation in specificity is much more apparent for reference translations of different numbers of sentences. As shown in Figure 6.2, a large portion of the differences are larger than those of two random sentences. Thus the usage of different *numbers* of sentences is a critical factor in changes in the specificity of reference translations.

	count		all		excl_multi		all, one-sent		token specificity	
length	H	¬H	H	¬H	H	¬H	H	¬H	H	¬H
0-10	0	15	-	0.0853	-	0.0853	-	0.0853	-	0.0077
10-20	27	548	0.2924	0.4131	0.4462	0.419	0.4697	0.4212	**0.0301**	0.0213
20-30	150	924	**0.5367**	0.7208	0.7739	0.7459	**0.8014**	0.7499	**0.0431**	0.0281
30-40	225	578	**0.7005**	0.8976	0.9242	0.9377	0.9407	0.9405	**0.0397**	0.0272
40-50	147	223	**0.7962**	0.948	0.9898	0.9915	0.9919	0.9916	**0.0353**	0.0229
50-60	70	44	**0.8456**	0.9593	0.9928	0.9992	0.9972	0.9993	**0.0301**	0.02
60-70	31	6	**0.8495**	0.9483	0.9998	0.9999	0.9997	0.9999	**0.0252**	0.0151
≥70	8	5	**0.8568**	0.9664	0.999	0.9999	0.9994	0.9999	**0.0196**	0.0141

Table 6.1: Average specificity of heavy (H) and non-heavy (¬H) reference translations given a length range of the source sentence: for all references (*all*), for only one-sentence translations (*excl_multi*), when all references are treated as one-sentence translations (*all, as-one-sent*), and the corresponding specificity per-token (*token specificity*). Bold font means significance ($p < 0.05$) when compared to non-heavy sentences. Content-heavy sentences have higher per-token specificity but translating them into multiple sentences leads to lower overall specificity.

6.2.2 Heavy vs. non-heavy sentences

To study the effect of a multi-sentence translation on specificity, we investigate the specificity of reference translations of heavy and non-heavy sentences. We group sentences according to the number of words in the source sentence to account for the correlation between sentence length and specificity. In Table 6.1 we show for each length range the average reference sentence specificity for (i) all reference translations for a Chinese sentence (column *all*) and (ii) only the one-sentence translations (column *excl_multi*).

By including multi-sentence translations, the specificity of the reference translations of heavy sentences drops significantly lower than non-heavy sentences. We study this decrease in specificity in terms of using multiple shorter sentences, and lexical variations involved in reorganizing the sentences. So to eliminate influence from sentence length and to separate the two causes, in column *all, one-sent*, we first run SPECITELLER for the reference translations as if all were single-sentence translations. This time heavy sentences are slightly higher in specificity (though mostly not statistically significant). Hence the drop in specificity is mostly due to the use of multiple shorter sentences itself.

Next, in Table 6.1, column *token specificity*, we show the per-token specificity, i.e., sentence specificity normalized by length, for heavy and non-heavy sentences. We use the *one-sent* setting for sentence specificity to control for the drop in specificity that is purely

pattern	specific-general	specific-specific	general-general	general-specific
%data	0.30	0.22	0.19	0.09
count	755	565	488	225

Table 6.2: Patterns of general and specific sentences within each multi-sentence translation among the heavy Chinese sentences. The least frequent pattern is general-specific, hence an INSTANTIATION relation is not likely to hold between two split components of a heavy sentence.

due to using multiple sentences, as described in the previous paragraph. Notably the token-level specificity is significantly higher among heavy sentences.

Therefore, when a translator uses multiple sentences to translate, the translations are made significantly less specific overall. At the same time, lexical alterations in this process lead to a significant increase in token-level specificity.

6.2.3 Intra-sentential specificity

Besides comparing specificity for heavy and non-heavy sentences and how a multi-sentence translation influences specificity, we now study how the content of a heavy sentence is split in terms of specificity. We focus on two-sentence translations since they account for 80% of the data. For each two-sentence translation of a heavy Chinese sentence, we check whether each split component is predicted general or specific. The counts for each pattern are shown in Table 6.2. In most cases a translator would split a sentence into segments that are of the same specificity category (specific-specific or general-general). The specific-general pattern—the translator processes most of the information before splitting the sentence—occurs less frequently. The least frequent case here is that a translator processes less information and leaves the rest to form a second, more specific sentence. The lack of general-specific pattern also indicates that it is less likely a translator split sentences due to an INSTANTIATION relation between the components.

6.3 Instantiation and other discourse relations

In this section we examine the occurrence of the INSTANTIATION relation among heavy and non-heavy sentences. We also study the distribution of discourse relations within the split components of a multi-sentence translation. To do this we consider the reference translations

	MTC+OpenMT					PDTB	
#words	Ref1	Ref2	Ref3	Ref4	Avg	Gold	Predicted
200	0.26	0.26	0.23	0.23	0.25	0.26	0.78
400	0.53	0.51	0.47	0.45	0.49	0.53	1.55
600	0.79	0.77	0.70	0.68	0.73	0.78	2.31
800	1.06	1.02	0.93	0.90	0.98	1.04	3.07
1000	1.32	1.28	1.16	1.12	1.22	1.30	3.86

Table 6.3: Number of sentence pairs with an implicit INSTANTIATION per 200 words in Chinese and English. The PDTB-Gold column shows the numbers for gold-standard annotations in the PDTB; others are predicted with the system in Section 3.2. The rate of predicted implicit INSTANTIATION is much lower in Chinese.

for each Chinese article separately, and detect implicit INSTANTIATION and explicit relations in the reference articles.

6.3.1 Implicit Instantiation in Chinese

Occurrence of Instantiation. We now study the rate of occurrance of the INSTANTIATION relation in Chinese and compare with that in English. We use the PDTB for comparison so that both datasets are from the news domain.

Using the NUS discourse parser (Lin et al., 2014), we detected 7.25 explicit INSTANTIATION relations across all reference translations among all the Chinese articles (3.5K sentences), a rate of about 0.2%. Compared to the rate of 0.67% in the PDTB[28], explicit INSTANTIATION is rare in both languages but even rarer in Chinese.

To account for implicit INSTANTIATION, we run our INSTANTIATION predictor described in Section 3.2 between the translations of each pair of source sentences in an article. Out of the 3,554 pairs of (source) sentences in all articles, there are 183, 173, 150, 146 implicit INSTANTIATIONs detected among the 4 reference translation articles (average 163, 4.6% of all sentence pairs).

In Table 6.3, we show the number of implicit INSTANTIATIONs per 200, 400, ..., 1000 words. Note that our predictor *overpredicts*, i.e., in the PDTB, the predicted INSTANTIATION rate is higher than the gold-standard rate. Hence the predicted rate in Chinese, shown in this table, should be higher than the actual rate of implicit INSTANTIATION. Still, the numbers here are much lower than the predicted case in English, indicating that compared

[28]We do a direct comparison here relying on the high accuracy of explicit discourse relation detection.

arg_1	mean	0.51	0.52	0.52	0.52	0.52
	std.dev	0.35	0.36	0.36	0.36	0.36
arg_2	mean	0.80	0.79	0.80	0.80	0.78
	std.dev	0.24	0.25	0.26	0.26	0.27

Table 6.4: Specificity of identified implicit INSTANTIATION arg_1s vs. arg_2s for 5 random draws among the 4 reference translations. The second arguments of identified INSTANTIATION are significantly less specific.

	Ref1	Ref2	Ref3	Ref4	Macro avg	Micro avg
Heavy&arg_1	6	7	7	5	-	-
%arg_1	3.28	4.05	4.67	3.42	3.60	3.85
%Heavy	0.68	0.78	0.79	0.56	0.70	0.70
Heavy&arg_2	106	105	90	88	-	-
%arg_2	57.92	60.69	60.00	60.27	59.67	59.72
%Heavy	11.94	11.82	10.14	9.91	10.95	10.95

Table 6.5: Counts of predicted INSTANTIATION arguments that are content-heavy, along with percentages of such sentences among all heavy sentences, among all arg_1s and among all arg_2s. There is a strong association between content-heavy sentences and arg_2s of INSTANTIATION.

to English, implicit INSTANTIATION is also rarer in Chinese.

Specificity among Instantiation arguments. We have shown that in English, the content in the first argument of an INSTANTIATION relation is general while the content in the second argument is specific. Here we verify that it is also the case in Chinese. For each identified implicit INSTANTIATION relation on the reference side, we compare the specificity scores of its arguments. For the same source sentence, if multiple references are identified as an argument, we randomly choose one of them. Table 6.4 shows that for each of the 5 random rounds, the specificity of arg_1s is consistently and significantly ($p = 0$) lower than arg_2s. This is a positive indicator that the general-specific arrangement of an INSTANTIATION relation is preserved across both English and Chinese.

6.3.2 Content-heavy sentences and Instantiation arguments

To gain insights of how content-heavy sentences are related to INSTANTIATION-like sentences and sentence pairs in Chinese, we see how often they appear among the arguments of an INSTANTIATION. In Table 6.5 we show the number of sentences that are both content-

	Inter-split		Intra-split	
	#	%	#	%
Comparison.Concession	0	0	8	**0.0118**
Comparison.Contrast	52.75	**0.0584**	37.75	0.0473
Contingency.Cause	17.25	0.0185	25.25	**0.0332**
Contingency.Condition	0	0	21	**0.0269**
Expansion.Alternative	1.75	0.0023	3	**0.0034**
Expansion.Conjunction	69.5	0.0765	99.75	**0.1267**
Expansion.Instantiation	2.75	**0.0036**	0	0
Expansion.Restatement	1	**0.0011**	0.25	0.0002
Temporal.Asynchronous	8	0.008	55.5	**0.0704**
Temporal.Synchrony	3.5	0.005	65.5	**0.0875**

Table 6.6: Average counts (#) and average percentages (%) of explicit discourse relations whose arguments are: between whole Chinese sentences, between split components within a sentence and within a split component. Larger percentages in each row are bolded. The distribution of discourse relations across split components is very different from that of relations that do not trigger a split.

heavy and predicted to be an arg_1 or arg_2 of INSTANTIATION, along with the percentages of such sentences among all content heavy sentences, all predicted arg_1s and all predicted arg_2s. Notably, heavy sentences account for about 60% of all identified INSTANTIATION arg_2s. This is substantially larger than the overall rate of heavy sentences. They are also very rarely predicted to be arg_1s. Furthermore, the rate of an INSTANTIATION arg_2 among content-heavy sentences is about 5% higher than in all sentences. Finally, if we consider the posterior probability of the INSTANTIATION classifier, the average across arg_1s is 0.33 while it is 0.48 for arg_2s. All of the above suggests a strong association between a content-heavy sentence and one that is similar to an INSTANTIATION arg_2 in English.

6.3.3 Relations across split components of heavy sentences

Previously we studied sentence specificity patterns when splitting a heavy Chinese sentence into multiple segments. Here we consider another signal that can be potentially involved to trigger splitting: explicit discourse connectives. We rely on the automatic classifier of Lin et al. (2014) to identify explicit discourse relations, since explicit relations can be reliably identified due to the presence of connectives (Pitler and Nenkova, 2009).

Specifically, we compare the distribution of *explicit* discourse relations within source sentences and between split segments of a source sentence. Table 6.6 shows, averaged across

	Ref1	Ref2	Ref3	Ref4	avg
Counts	17	26	7	16	16.5
% heavy	3.1	4.4	1.7	3.0	3.0

Table 6.7: Number of predicted implicit INSTANTIATION relations within multi-sentence translations, along with its percentages among heavy sentences, for each translator separately (*Ref1-4*) and averaged among all translators (*avg*). Implicit INSTANTIATION is not a strong trigger for splitting heavy sentences.

references, discourse relations whose arguments: are within the same source sentence but in different English sentences after splitting (count and percentage over split sentences); are within the same source sentence and within the same English sentence even after splitting (count and percentage over split sentences). Numbers largest in each *row* are bolded. The distributions of relations are clearly different. For example, the explicit CONTRAST, INSTANTIATION and RESTATEMENT relations are more involved across split segments while relations such as CONJUNCTION, TEMPORAL, CONDITION and CONCESSION are less likely to trigger a split.

Finally, we show the counts for *implicit* INSTANTIATIONS, along with the percentage of heavy sentences these counts account for, in Table 6.7. On average, only 3% of the heavy sentences contain an implicit INSTANTIATION, which is lower than the overall rate of INSTANTIATION in Chinese (4.6%). Once again we have a positive indication that the *implicit* INSTANTIATION relation is not a strong signal for a splitting point.

6.4 Predicting heaviness

So far we have shown that there are significant differences in sentence- and token-level specificity between heavy and non-heavy sentences, and that sentences similar to the second argument of an English INSTANTIATION relation are more likely to be heavy. We now investigate whether these two aspects are complementary to sentence length and the syntactic characteristics laid out in Section 5.2. To do this, we predict whether a source sentence is content-heavy using the following features in a logistic regression model:

- **P**: Posterior probability from our model of content-heavy sentence prediction described in Section 5.2. Recall that the features in this model are syntactic and that

Features	Accuracy	Precision	Recall	F
L+P	0.7780	0.5759	0.7198	0.6399
L+P+I	0.7792	0.5774	0.7243	0.6426
L+P+ST	0.7817	0.5821	0.7209	0.6441
L+P+S	0.8118	0.6308	0.7548	0.6892
L+P+ST+S	**0.8139**	**0.6347**	0.7559	**0.6900**
all	0.8118	0.6303	**0.7571**	0.6879

Table 6.8: Accuracy, precision, recall and F-measure for content heavy sentence prediction. *P*: posterior probability from our content-heavy classifier in Section 5.2; *L*: sentence length; *S*: specificity; *ST*: per-token specificity; *I*: probability of being a predicted INSTANTIATION arg_2. Specificity information complements L and P in predicting heavy sentences.

they are extracted from the source sentence.

- **L**: Length of the source sentence.
- **S**: Specificity (SPECITELLER score) of a reference translation. Since parallel corpora usually contain one reference translation per source sentence, we randomly sample one reference translation for each source sentence.
- **ST**: Per-token specificity of the randomly sampled reference translation as discussed in Section 6.2.2.
- **I**: Posterior probability of the sentence being an arg_2 of an implicit INSTANTIATION.

Performances of 10-fold cross validation are shown in Table 6.8. Each additional feature related to specificity is added on top of length (L) and the syntactically-derived probability (P). We see that specificity of the reference translations (S) gives the largest margin of improvement; per-token specificity (ST) also complements the overall specificity in all metrics. The best system is *L+P+ST+S*, showing that specificity information captures additional information not present in length and syntactically derived probability. The probability of being an arg_2 of INSTANTIATION (I) complements L+P; however when we add it on top of specificity (*all*), the performance went down. Hence while most INSTANTIATION arg_2s are heavy sentences, their low fraction among heavy sentences makes them uninformative when identifying heavy sentences.

6.5 Conclusion

We further study content-heavy sentences in terms of different ways general and specific information is expressed. We found that while these sentences involve higher specificity per-token, by translating a sentence into multiple sentences, the overall specificity is lower. Lowering the specificity of translated text likely increases its intelligibility, as they are indicators for sentence simplification 7.1. Heavy sentences also demonstrate a strong association with the implicit INSTANTIATION relation. Both of these factors are complementary to previously investigated length and syntactic factors. Finally, we observe distinct discourse relation distributions that have arguments across split segments of a sentence, further signaling different ways of packaging content in Chinese and English.

Chapter 7

Coherence preferences: monolingual

Successful communication is a careful balancing act: speakers must gauge the appropriateness of producing statements that are highly detailed vs. statements that are more general. Whereas too much detail may overwhelm the listener, vague statements with too little detail can sound vacuous. In this chapter we investigate, within the same language, personal characteristics that potentially contribute to readers' perception and production of text specificity.

First, we consider beginner readers by studying simplified sentences in reading material created for child-appropriate reading levels and learners of English. We find that the specificity of simplified sentences written for these readers is significantly lower than their original versions. Sentences with high specificity are also more likely to be simplified than others.

We then consider gender and adults with varying autism-like symptoms. We conduct a pilot study targeting both the perception of text specificity and the specificity of text written by the subjects. We find that male and female subjects differ significantly in specificity perception, and weak, non-statistically significant trends that people with more autism-like

Content in Section 7.1 is published at AAAI 2015 (Li and Nenkova, 2015b). Content in Section 7.2 is presented as a poster at IMFAR 2017 (Li et al., 2017b).

symptoms perceive text to be more general and write more general text than others. We discuss the limitations of our study and lay out future directions.

7.1 Sentence specificity and text simplification

To investigate the relationship of sentence specificity and reading ability we discuss the role of sentence specificity in text simplification applications. Specifically we wish to quantify the extent to which specificity changes during sentence simplification and to determine if sentence specificity is a useful factor for determining if a sentence needs to be simplified in the first place.

To give context to our findings, we also analyze the relationship between simplification and sentence length, automated readability index (ARI)[30] and language model perplexity[31]. We carry out analysis on two aligned corpora: Simple Wikipedia/Wikipedia and Britannica Elementary/Encyclopedia Britannica.

The Wikipedia corpus (Kauchak, 2013) is created for children and people who are learning English. It features automatic aligned sentence pairs from the Simple Wikipedia and the original English Wikipedia. The dataset consists of 167,689 aligned pairs, among which about 50K are the same sentences across Simple and original Wikipedia.

The Britannica corpus is constructed by (Barzilay and Elhadad, 2003), featuring the Britannica Elementary. Reading material in Britannica Elementary is especially written to be appropriate for children's reading level. People were asked to align sentences that share semantic content from several articles in the Britannica Elementary and the original Encyclopedia Britannica. There is only one pair where the two sentences are the same.

7.1.1 Specificity as simplification objective

First, we studied the extent to which simple and original texts in the two corpora vary in terms of their automatically predicted specificity by SPECITELLER. We contrast these with

[30] We also considered Kincaid, Coleman-Liau, Flesh Reading Ease, Gunning Fog Index, LIX, SMOG and RIX. ARI was the readability measure that showed biggest difference in readability between original and simplified sentences.

[31] Our language model is trained on the New York Times articles from 2006. It is a trigram model using Good-Turing discounting, generated by SRILM (Stolcke, 2002).

		%pairs	mean, simplified	mean, original
Wikipedia	ARI	73.60	9.76	12.94
	specificity	70.86	0.57	0.70
	perplexity	62.99	1272.61	1539.48
	length	55.19	23.74	27.57
Britannica	ARI	82.14	8.82	14.13
	specificity	77.12	0.45	0.70
	perplexity	74.29	635.50	1038.36
	length	73.42	19.75	30.10

Table 7.1: Percentages of original-simplified sentence pairs with lower attribute values for the simplified side (%pairs), along with mean values for each attribute among simplified and original sentences. Specificity of simplified sentences are remarkably lower.

the differences in average sentence length, average sentence readability and perplexity. For both corpora, we excluded pairs where the simplified version and the original are identical.

For both corpora, there can be more than one sentence on each side of an aligned pair. So to measure specificity, we first classify each sentence in each pair of the corpora using the final combined classifier obtained from co-training. Following the definition in Louis and Nenkova (2011a), the specificity of side $i \in$ {simplified,original} of a pair p is calculated as:

$$spec(p_i) = \frac{1}{\sum_{s \in p_i} |s|} \sum_{s \in p_i} |s| \times Pr(\text{specific}|s) \tag{7.1}$$

Here s denotes a sentence in p_i, $|s|$ denotes the length of the sentence and $Pr(\text{specific}|s)$ denotes the posterior probability of the classifier assigning sentence s as specific.

In Table 7.1, we show the average value of the attributes for simplified and original sides. For all attributes, we observe a significant ($p < 0.01$) drop in their values for the simplified sentences. More importantly shown in Table 7.1 are the percentage of pairs for each attribute where the simplified side has a lower value than the original side. The higher the percentage, the more one would expect that the attribute needs to be explicitly manipulated in a procedure for sentence simplification. Not surprisingly, the highest value here is for ARI, as improved readability is the goal of simplifying sentences for junior readers. Specificity score closely follow ARI, with about 71% and 77% of the simplified sentences showing lower specificity in the Wikipedia and Britannica corpora respectively. The numbers are much higher than those for sentence length and perplexity.

attribute	Wikipedia	Britannica
ARI	0.6158	0.7019
specificity	0.6144	0.6923
length	0.5454	0.6154
perplexity	0.3966	0.3308

Table 7.2: Precision for identifying sentences to simplify. Specificity and ARI outperform other attributes.

attribute A	attribute B	Wikipedia	Britannica
length	ARI	0.7897	0.7822
specificity	length	0.6996	0.7669
specificity	ARI	0.5975	0.6788
specificity	perplexity	0.3695	0.5306
perplexity	ARI	0.2454	0.3597
length	perplexity	0.1073	0.2293

Table 7.3: Spearman correlation for the attributes in original sentences. The correlation between specificity and ARI are not very high.

7.1.2 Identifying simplification targets

Now we analyze if specificity is an indicator that an *individual* sentence should be simplified in the first place. We train a predictor to detect a sentence that needs to be simplified with each of the sentence attributes in turn. Our positive training examples are those original sentences that have been simplified. All other sentences, including all of the examples where the simple and the original sentences are the same, serve as negative examples. We report the precision of each single-attribute classifier in identifying sentences in the original data that need to be simplified.

In Table 7.2 we show for each attribute, the precision for identifying sentences that need to be simplified, obtained by logistic regression via 10-fold cross-validation[32]. We also record in Table 7.3 the Spearman correlation between the attributes. For both corpora, sentence specificity is the second best attribute, closely following ARI with less than 1% difference in precision. Sentence length itself is not that good to identify which sentences require simplification. Perplexity from language model is the least helpful for this task. The correlation between specificity and ARI are not very high, indicating that the two attributes complement each other, each being useful as an indicator.

[32]We downsampled the negative class for the Wikipedia corpus such that the positive and negative classes are of the same size.

7.1.3 Conclusion

We analyze the impact of sentence specificity on sentence simplification and showed that sentence specificity is not only a useful objective for simplification, but also indicative in identifying sentences that need simplification. Hence packing too much detail in text may not be desirable for simplified text targeting beginner readers. Further, future applications can use specificity as an indicator to detect such sentences to help these target audiences.

7.2 Gender, Autism Quotient scores and the perception and production of text specificity

In this pilot study, we begin to test whether group differences are associated with the perception and production of text specificity. We start with two aspects: gender and Autism Spectrum Disorder tendencies in typical adults.

Autism Spectrum Disorder (ASD) is a neurological and developmental disorder that impacts an individual's communication and social abilities. It is a "spectrum" as its symptoms affect individuals to different degrees. According to the Center for Disease Control, about 1 in every 68 births in the US suffer from the disorder, and that this number has increased by 123% from 2002 to 2010[33]. The cost associated with the disorder is estimated to be about $2.4 million per person, and that the cost is higher for adults than for children (Buescher et al., 2014). One of the most notable symptoms of the disorder is that an individual may have "a lasting, intense interest in certain topics, such as numbers, details, or facts"[34]. Hence we would like to explore whether some communication difficulties experienced by individuals with ASD may be due to challenges in understanding or producing appropriate levels of communicative specificity during conversation. Our long-term goal is to determine the extent to which existing automated tools and computational theories of language vagueness and specificity can elucidate differences between individuals with ASD and typical controls, for the purposes of enhancing screening, diagnosis, treatment planning, and intervention response measurement.

[33]Center for Disease Control, https://www.cdc.gov/ncbddd/autism/addm.html

[34]National Institute of Mental Health, https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

We are also motivated by language variation among different demographic groups. Characterizing language variation across gender is an established and important aspect of sociolinguistics (Coates, 2015). It can also help user attribute prediction (Burger et al., 2011; Flekova and Gurevych, 2013; Bamman et al., 2014; Volkova and Yarowsky, 2014; Johannsen et al., 2015; Preoţiuc-Pietro et al., 2016) and addressing gender-related biases in text (Volkova et al., 2013; Hovy, 2015; Flekova et al., 2016). Most of these work explore gender and language through lexical usage and syntax, especially via social network data such as Twitter and user reviews. Our work complements existing studies by considering the perception and production of text specificity—as a property at the sentence level and above—on news data.

We uncover significant differences in text specificity perception between male and female subjects, which clearly indicates that specificity is an aspect worth considering in sociolinguistic research. On the other hand, we found a weak association between more ASD-like symptoms and an increase in perceived text specificity that is not statistically significant. This negative result suggests limitations of our current approach and improvements in strategy in future studies.

7.2.1 Overview

We designed a single-session user study. **To assess ASD traits**, we gather each subject's score for the Autism-Spectrum Quotient (AQ) test (Baron-Cohen et al., 2001). This test was developed as a quick screening tool that can be administered with ease. In contrast, usual diagnostic tools like the Autism Diagnostic Observation Schedule would require a trained psychologist and take up to 60 minutes to administer with the subject. The AQ test involves 50 questions such as the ones below:

1. I prefer to do things with others rather than on my own.
2. I prefer to do things the same way over and over again.
3. If I try to imagine something, I find it very easy to create a picture in my mind.

The answers can be "definitely agree", "slightly agree", "slightly disagree" and "definitely disagree"; a score 1 is added if the answer indicates autism traits (e.g., "agree" for 2). Baron-Cohen et al. (2001) did not distinguish between "definitely" and "slightly". The

maximum possible score is 50; their study suggested a score above 32 to indicate clinical diagnosis of ASD.

We also collect information regarding the subject's **gender and native language**. **For specificity perception**, we follow Section 4.1 and ask each subject to rate the specificity of a sentence on a scale of 0 (most specific)—6 (most general). In this study we re-use the 40 sentences with the highest agreement from the New York Times articles used for annotation in Section 4.1. **For specificity in text production**, each subject reads through two full New York Times articles and are asked to write a summary for each text. We selected one text that contains more details (specificity 0.77, 1981 words) and another of average specificity (specificity 0.48, 5231 words)[35].

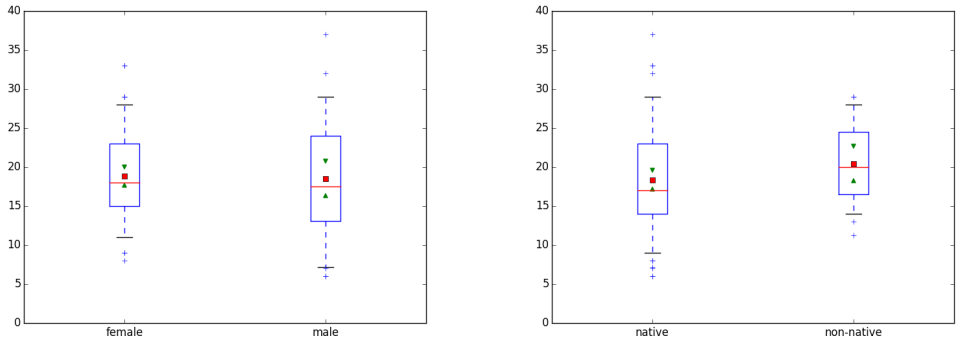
7.2.2 Subjects

Recruitment and filtering. We recruited a total of 144 subjects among undergraduate students at the University of Pennsylvania. We exclude the data from three groups of subjects: (a) subjects who are in a completely different age group as others; (b) subjects who wrote summaries in less than 30 seconds, as they likely were not paying attention to the task; and (c) subjects whose specificity ratings do not correlate with the average of others', as long as this (non-)correlation is not related to the AQ scores. They also were likely not paying attention to the task.

To determine (c), for each subject $S_i \in \{Subj\}$ we calculate the Spearman correlation between the specificity rating for each sentence given by S_i and the average of $\{Subj\} \setminus S_i$. We found one subject whose rating negatively correlate with others (-0.48) and 11 subjects whose correlation values are not significant ($p > 0.05$, all correlation values are less than 0.3). The AQ values of these 12 subjects do not significantly differ from the others, so we exclude them from all analysis.

Gender and native language. We gather two pieces of information from our subjects that we believe would be important to consider: gender and native language. Prior work has shown that both traits are associated with numerous differences in writing styles (Flekova et

[35]The overall specificity of these articles are automatically assessed using Speciteller (Section 3.3.) with a score between 0 (most general) to 1 (most specific)



(a) Female vs. male.

(b) Native vs. non-native speakers.

Figure 7.1: Boxplots of AQ scores among subject groups. Red line: median; red dot: mean; green arrows: 95% confidence intervals; boxes/whiskers: 75%-95% quantiles. AQ scores do not differ significantly across gender but differs for non-native speakers.

al., 2016). Here we look for their influence on text specificity. Meanwhile, we check whether they are significant factors in AQ differences so that we can take them into account in our AQ-related analysis.

We conduct a Wilcoxon rank sum test to see whether the distribution of AQ scores among male and female subjects are significantly different. Figure 7.1a shows AQ scores among male and female subjects. Female subjects are more consistent in AQ than male subjects. There is no significant difference ($p = 0.83$) between the two groups.

Figure 7.1b shows AQ scores for native and non-native English speakers. Non-native speakers have almost significantly higher ($p = 0.054$) AQ scores than native speakers. Hence we exclude non-native speakers from all analysis related to AQ.

Finalized subject pool. To compare native and non-native speakers, we consider only female subjects (85 subjects) as there are only two male subjects who are non-native speakers. When analyzing gender and AQ, we exclude non-native speakers to minimize bias in language.

7.2.3 Specificity perception

We use 40 sentences from our annotation in Section 4.1 as stimuli for specificity rating. The sentences satisfy two criteria: (a) the standard deviation of the ratings from our three

Gender	# subj	95% intv		mean	std.dev
male	44	0.5964	0.8317	0.7141	0.3631
female	64	0.3727	0.6251	0.4989	0.3896

Table 7.4: Numbers of male and female subjects, 95% confidence intervals and mean/standard deviation of specificity perception values. Male subjects rate sentences significantly less specific than female subjects.

annotators are less than 1; (b) they are sampled according to the overall distribution of sentence specificity ratings from Section 4.1.

To have a fair assessment of sentence specificity perception that is independent of the specificity of the sentences themselves, we look at for each sentence a subject's *deviation* of specificity rating from average ratings collected from raters unrelated to the study. We call this the *perception value*. We collect these average ratings from Mechanical Turk workers. Just as with the subjects, we exclude workers whose rating correlation with everyone else's is less than 0.3. This results in about 30 ratings per sentence. Their average values significantly correlate with ratings from our trained annotators in Section 4.1 (Spearman correlation 0.76).

Gender. We first check whether the perception of specificity differs among male and female subjects. Table 7.4 shows the 95% confidence interval, average and standard deviation of perception values in each group. Female subjects tend to rate sentences more specific than male subjects. The differences are significant (Wilcoxon sign rank $p < 0.05$).

AQ scores. We analyze male and female subjects separately due to their difference in specificity perception. We partition subjects into a control group (Group 1) and a high-AQ group (Group 2) for each gender. Ideally subjects in Group 2 should have AQ scores above 32 (indicating clinical diagnosis), but we have only one subject with a score in this range. So we resort to choosing another cutoff. In their original study, Baron-Cohen et al. (2001) listed for each AQ score the number of ASD and control subjects above the score. We select an AQ score cutoff for each gender such that: (a) we have at least 10 subjects in each group; (b) in Baron-Cohen et al. (2001), the number of control subjects above the cutoff is less than 15%. Table 7.5 shows the cutoffs and the number of subjects in each group. The distribution of AQ values, along with the cutoffs, are shown in Figure 7.2.

	Cutoff	#subj (G1)	#subj (G2)
Male	24	33	11
Female	23	53	11

Table 7.5: Cutoff AQ scores for the high-AQ group and number of subjects for control (G1) and high-AQ (G2) groups.

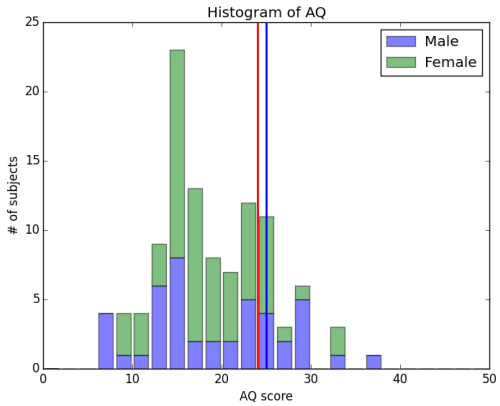


Figure 7.2: Distribution of AQ scores. Vertical lines indicate cutoff AQ scores of the control and high-AQ groups (red: female, blue: male).

	Rating			Out of interval	
	Corr.	G1	G2	G1	G2
Male	0.01	0.736	0.649	0.756	0.757
Female	-0.10	0.500	0.493	0.736	0.766

Table 7.6: Spearman correlations of specificity perception values and AQ scores, average perception values among subjects for each group, and fractions of raw sentence ratings outside 95% confidence intervals of crowd ratings. There is a non-significant trend that subjects with higher AQ scores give more specific ratings.

To look for relationships between perception values and AQ scores, we consider three measures tabulated in Table 7.6:

1. Spearman correlation between perception values and AQ scores;
2. Whether the perception values are different among subjects in the control group (G1) vs. the high-AQ group (G2);
3. Whether the percentage of raw sentence specificity ratings that are outside of 95% confidence interval from MTurk ratings are different among subjects in the control group (G1) vs. the high-AQ group (G2).

	Gender	# subj	95% intv		mean	std.dev
Article 1	male	44	0.4241	0.5591	0.4916	0.2195
(spec 0.77)	female	64	0.3967	0.5004	0.4486	0.2061
Article 2	male	44	0.3516	0.4812	0.4164	0.2016
(spec 0.48)	female	64	0.3598	0.4631	0.4115	0.2052

Table 7.7: Numbers of male and female subjects, 95% confidence intervals and mean/standard deviation of summary specificity. There is a non-significant trend that female subjects tend to write less detailed summaries.

	Article 1 (spec 0.77)			Article 2 (spec 0.48)		
	Corr.	G1	G2	Corr.	G1	G2
Male	-0.022	0.500	0.465	-0.14	0.430	0.376
Female	0.088	0.443	0.477	-0.10	0.420	0.371

Table 7.8: Spearman correlation of summary specificity and AQ scores, average summary specificity ratings among subjects for each group. There is a non-significant trend for Article 2 that subjects with higher AQ scores tend to write less detailed summaries.

Subjects in the high AQ group in general give more specific ratings. Among females in the high AQ group, the fraction of ratings outside of 95% confidence intervals of crowd ratings is larger than the control group, so the deviation from how specific these sentences are normally perceived is larger. However, these trends are not statistically significant.

7.2.4 Specificity of text produced

Gender. We now look for the relationship between the specificity of summaries written by the subjects and their gender. Here the overall per-word specificity of the summaries is measured by SPECITELLER. The scales are from 0 (most general) to 1 (most specific). Table 7.7 illustrates that female subjects tend to write summaries that are less detailed than male subjects; however this tendency is not statistically significant.

AQ scores. Finally, we explore the relationship between the specificity of summaries written by the subjects and their AQ scores. Table 7.8 shows for each article, the Spearman correlation as well as the average specificity ratings for G1 and G2.

For the less specific article (Article 2), both male and female subjects with higher AQ scores tend to write more general summaries. However this tendency to write more general summaries is not statistically significant. We have observed no trend for the more detailed

article (Article 1).

7.2.5 Discussion and conclusion

In this pilot study we start to investigate potential connections between text specificity perception and production, and group characteristics including gender and ASD traits. We assess both the perception of sentence-level specificity as well as the specificity of summaries generated by the subjects. There are significant differences in specificity perception among male and female subjects. Also, there are non-significant trends that subjects with higher AQ scores may have a tendency to perceive sentences more specific than others and to write more general summaries.

One major limitation of the study is in terms of subjects. Currently, we have only one subject whose AQ scores are above the recommended cutoff for confirming diagnosis in the AQ test. Even with the reduced cutoffs, the number of subjects above them is small and many of them have AQ scores close to the cutoff. In addition, we have fewer male subjects than female. One obvious future direction is to recruit subjects with ASD diagnosis, as well as a balancing number of male and female subjects. Second, the current stimuli are sentences and articles from the New York Times, so one way to extend to broader populations is to find child-appropriate or lower-IQ appropriate stimuli. Third, since the sentences in the perception analysis are selected according to the overall specificity distribution from our previous study, there are naturally fewer sentences that are very general or very specific. We will include more sentences at either ends of the specificity spectrum in future work. Finally, for the summary analysis, we have only studied the overall per-word specificity in the summaries. A great way to carry out further analysis is to investigate how the elements and concepts are selected and expressed within the summaries.

Our study is the first to examine the perception and production of linguistic specificity in people with varying degrees of autism-like symptoms. Although preliminary, our results reveal previously unexplored gender differences, which we plan to confirm with a larger sample and participants with official ASD diagnoses. This is the first step in a research program aimed at developing technology-augmented interventions to visualize the mismatch in expectations regarding specificity to both typical people and individuals with ASD, to

improve the quality of communication across the two groups. Furthermore, we aim to develop a battery of tests to quantify an individual's perception of specificity, and explore the use of these metrics for assessing intervention effectiveness.

7.3 Discussion and future work

We examine non-linguistic aspects that affect communication: reading ability, gender and traits for ASD. Our results show that lower specificity associates strongly with lower reading ability; female readers perceive sentences to be more specific than male. Also, our pilot analysis leads to weak observations that adults with more ASD-like symptoms may tend to rate sentences as more specific and may write less detailed text.

These analyses reveal that how much detail to express in text and how to organize them is associated with the style of written communication and comprehension. In future work, we would like to refine these analyses, e.g., to consider multiple reading levels, recruit broader subjects and make our stimuli broadly accessible. We would also like to consider specificity changes in discourse. Finally, we seek to explore other group characteristics.

Work in this chapter can further lead to a broad future direction in tailoring text to a level of details suitable for the given target reader group. As first steps, we can start with text simplification and summarization systems and take into account the specificity of original and generated text.

Chapter 8

Conclusion

This thesis investigates preferences of the organization of general and specific content among different groups of readers. We propose novel methods to quantify text specificity drawn from insights in discourse structure. We show that conventions and expectations of text specificity vary in cross-lingual context and among audiences with different gender and reading abilities. We further identify factors in such variance that when not addressed, impact text coherence and intelligibility.

8.1 Summary of contributions

Discourse organization and text specificity. We characterize general and specific content in text with the discourse relation INSTANTIATION and text specificity. INSTANTIATION is the most prominent discourse relation related to the change of level of details in text, and is mostly implicit. By studying the characteristics of INSTANTIATION, we substantially improve the identification of this relation (Li and Nenkova, 2016) (Section 3.2).

We propose new annotation guidelines for sentential and subsentential specificity that quantify: level of sentence specificity, expressions lacking in specificity within the sentence as well as discourse effects associated with those expressions (Li et al., 2016) (Section 4.1). Analyses from our pilot corpus show that the lack of specificity, especially when not elaborated in prior context, triggers high-level text comprehension questions among readers.

We build systems to quantify the specificity of sentences and words. With INSTANTI-

ATION as seed training data, we design and publicly release a semi-supervised system to obtain sentence specificity with speed and accuracy (Li and Nenkova, 2015b) (Section 3.3). This system utilizes only string surface information, so that it can be easily adapted to other applications for analysis. Further, we predict words in a sentence that may need elaboration using an attention network to predict sentence level specificity (Section 4.2). Our method outperforms one that derives word specificity independently from the sentence itself. This is the first work towards automatically quantifying subsentential specificity.

Coherence preferences in cross-lingual context. We identify discourse devices that significantly impact machine translation quality (Li et al., 2014) (Section 5.1). We show that the amount of information conventionally packaged in a Chinese sentence is different from that in English and this difference highly impacts the quality of translated text for both human and machine. We define these content-heavy sentences from translated text of multiple translators and preferences from readers. These sentences are more specific than non-heavy sentences; by using multiple sentences to translate, the specificity of translated text is lowered. We present a high performing system to detect these sentences. Our method is able to identify a set of sentences much more difficult for machines to translate (Li and Nenkova, 2015a) (Section 5.2).

Coherence preferences and group characteristics. We discover that preferences for text specificity vary among broad groups of readers: people with different gender, reading ability, and autism traits. We show that sentence specificity is an important characteristic in simplified text. Furthermore, sentences deemed specific is an effective indicator that it should be simplified for beginner readers (Li and Nenkova, 2015b) (Section 7.1). This aspect is complementary to existing metrics of readability such as the Automated Readability Index (Senter and Smith, 1967).

Our pilot study uncovers clear differences in the perception of text specificity among male and female readers. In addition, we conduct the first analyses of text specificity perception and production among people with varying autism-like symptoms (Section **??**). Although the study only discovers weak trends, it shows clear future directions that we plan to undertake.

8.2 Future work

In this final section, we summarize future directions led by our work.

The INSTANTIATION relation. One aspect that we haven't explored when characterizing INSTANTIATION (Section 3.2) is textual entailment. As pointed out in Section 3.2.4, even though in theory the second argument of INSTANTIATION should entail the first, few instances of the relation in the PDTB are automatically recognized as entailment. We found that the entailment relationship appears to be at phrase or clause levels, and often depends on context and external knowledge. We believe future work exploring these directions can benefit RTE systems and INSTANTIATION recognition.

When comparing INSTANTIATION with SPECIFICATION, we pointed out that the *change* in specificity across SPECIFICATION arguments may not be as large as that in INSTANTIATION, especially if the first argument of SPECIFICATION does not need to be particularly general or the second argument particularly specific (Section 3.4). This hypothesis, if confirmed, will bring new insight into discourse relations and specificity. We leave for future work to have a fair judgement of this hypothesis, using a measure of specificity independent of either relations (e.g., via human judgements, such as that outlined in Section 4.1).

Subsentential specificity. We present an annotation guideline and a pilot corpus to annotate the degree of sentence specificity, and the cause and effect of underspecified text (Section 4.1). In this annotation, we did not separate if an underspecified text segment is elaborated in upcoming context or not in upcoming context. We also leave for future work to analyze the content of the underspecified segments and their associated questions, which can be useful for gaining further insights into what needs elaboration and what causes vagueness.

Our work proposes the first model to predict underspecified words within a sentence (Section 4.2). As pointed out in Section 4.3, there are multiple ways to improve the model, for example, to train on more data so that more powerful models can be adopted, to gain sharper attention weights using repeated attention, and to obtain structure on top of the current token-level prediction with structured attention. Future work can also tackle the

prediction of the number of underspecified tokens, along with how and where they can be resolved.

Cross-lingual analysis. In Section 5.1, we pointed out that both Arabic and Chinese have sentences that need multiple English sentences to translate. However we did not find these sentences to be especially problematic for Arabic-English translation. Future work can explore this negative result, and uncover linguistic constructs that lead to this contrasting finding between Arabic and Chinese. Future work can also look into more languages, especially those with more extreme differences in punctuation usage (e.g., Thai).

To identify content-heavy sentences in Chinese which need multiple English sentences to translate, we developed a system with rich syntactic features. We also pointed out differences in discourse relation distribution across split components in a heavy sentence vs. those not involved in splitting (Section 6.3.3). As pointed out in Section 5.3, one obvious future direction is to incorporate the insight from our work to improve Chinese to English machine translation.

In terms of specificity, we discovered strong associations between content-heavy Chinese sentences, text specificity and the second argument of INSTANTIATION (Chapter 6). Future work can further explore specificity across different languages, e.g., sentence specificity prediction in Chinese.

Specificity and sentence simplification. Section 7.1 shows strong associations between specificity and simplified sentences. When characterizing sentences that need simplification, specificity is as indicative as and complementary to readability. Furthermore, we found that often the simplified version of a sentence uses multiple sentences to express the content in the original. We leave to future work to incorporate specificity and our insights in content-heavy sentences into sentence simplification systems.

Specificity and demographics. Section 7.2 presents our pilot study exploring specificity perception variation across varying autism symptoms. While our study did not lead to statistically significant findings, we pointed out several ways to improve the experiment: recruiting subjects with clinically diagnosed ASD, expanding the applicability of our stimuli

and extending our analysis on subject-produced summaries.

Our work exploring links between specificity and gender, reading abilities and autism symptoms opens new directions for future work to go further into aspects in socio-demographics and personal background (Section 7.3). Research in social media text has found distinctive language usage across people of different genders, income levels, personality and political views. We leave for future work to investigate how specificity is perceived and organized when these aspects vary.

References

Manish Agarwal, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. 2001. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, 31(1):5–17.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

David I Beaver and Brady Z Clark. 2009. *Sense and sensitivity: How focus determines meaning*, volume 12. John Wiley & Sons.

Daniel Beck, Kashif Shah, and Lucia Specia. 2014. SHEF-Lite 2.0: Sparse multi-task gaussian processes for translation quality estimation. In *Ninth Workshop on Statistical Machine Translation*.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing*, September.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 69–73.

Or Biran and Kathleen McKeown. 2015. PDTB discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2201–2211.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

AS Buescher, Cidav. Z, M Knapp, and DS Mandell. 2014. Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA Pediatrics*, 168(8):721–728.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54.

Gregory N. Carlson. 2005. Generics, habituals and iteratives. In *Encyclopedia of Language and Linguistics*. Elsevier.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72.

Joyce Y. Chai and Rong Jin. 2004. Discourse structure for context question answering. In *HLT-NAACL 2004: Workshop on Pragmatics of Question Answering*, pages 23–30.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1041–1044.

Pi-Chuan Chang, Daniel Jurafsky, and Christopher D. Manning. 2009a. Disambiguating "DE" for Chinese-English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009b. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*.

Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540.

Ian Palmer Cook. 2016. *Content and Context: Three Essays on Information in Politics*. Ph.D. thesis, University of Pittsburgh.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Osten Dahl. 1975. On generics. *Formal Semantics of Natural Language*, pages 99–111.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was it good? It was provocative." Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172.

Peter Dixon. 1982. Plans and written directions for complex tasks. *Journal of Verbal Learning and Verbal Behavior*, 21(1):70–84.

Peter Dixon. 1987. The processing of organizational and component step information in written directions. *Journal of memory and language*, 26(1):24–35.

Alex Djalali, David Clausen, Sven Lauer, Karl Schultz, and Christopher Potts. 2011. Modeling expert effects and common ground using questions under discussion. In *AAAI Fall Symposium: Building Representations of Common Ground with Intelligent Agents*.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *Transactions of the Association for Computational Linguistics*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008.

N. Elhadad, M.-Y. Kan, J.L. Klavans, and K.R. McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33(2):179 – 198. Information Extraction and Summarization from Medical Documents.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Federico Fancellu and Bonnie Webber. 2014. Applying the semantics of negation to SMT through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 511–521.

Lucie Flekova and Iryna Gurevych. 2013. Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany, August. Association for Computational Linguistics.

Austin F Frank and T Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the cognitive science society*, volume 30.

William Frawley. 1992. *Linguistic Semantics*. L. Erlbaum Associates.

Lyn Frazier, Charles Clifton Jr., and Britta Stolterfoht. 2008. Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, 106(1):299 – 324.

Annemarie Friedrich and Manfred Pinkal. 2015. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1272–1281.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 273–280.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613.

David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. In *Linguistic Data Consortium*.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

NIST Multimodal Information Group. 2013. NIST 2008-2012 Open Machine Translation (OpenMT) Progress Test Sets LDC2013T07. In *Linguistic Data Consortium*.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Short Papers*, pages 49–52.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pages 299–305.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL 2004: Main Proceedings*, pages 185–192.

Donald Hindle. 1983. Discourse organization in speech and writing. In Muffy E. A. Siegel and Toby Olson, editors, *Writing Talks*. Boynton/Cook.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November.

Alexander Hogenboom, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Communication of the ACM*, 58(7):69–77, June.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762.

Shudong Huang, David Graff, and George Doddington. 2002. Multiple-Translation Chinese Corpus LDC2002T01. In *Linguistic Data Consortium*.

Shudong Huang, David Graff, Kevin Walker, David Miller, Xiaoyi Ma, Christopher Cieri, and George Doddington. 2003. Multiple-Translation Chinese (MTC) Part 2 LDC2003T17. In *Linguistic Data Consortium*.

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 13–24.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342.

Yaohong Jin and Zhiying Liu. 2010. Improving Chinese-English patent machine translation using sentence segmentation. In *International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–6.

Meixun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2004. Segmentation of Chinese long sentences using commas. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*, pages 1–8.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Andrew Kehler. 2004. Discourse coherence. *The handbook of pragmatics*, pages 241–265.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *International Conference on Learning Representations*.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387.

Alex Lascarides and Nicholas Asher, 2007. *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*, pages 87–124. Springer Netherlands.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Junyi Jessy Li and Ani Nenkova. 2014. Reducing sparsity improves the recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 199–207.

Junyi Jessy Li and Ani Nenkova. 2015a. Detecting content-heavy sentences: A cross-language case study. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1271–1281.

Junyi Jessy Li and Ani Nenkova. 2015b. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*.

Junyi Jessy Li and Ani Nenkova. 2016. The instantiation discourse relation: A corpus analysis of its properties and improved detection. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. Data Distillation for Controlling Specificity in Dialogue Generation. *ArXiv e-prints*, February.

Junyi Jessy Li, Julia Parish-Morris, Leila Bateman, and Ani Nenkova. 2017b. Autism quotient scores modulate the perception and production of text specificity in adult females. In *The International Meeting for Autism Research*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 4.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233.

Annie Louis and Ani Nenkova. 2010. Creating local coherence: An empirical assessment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 313–316.

Annie Louis and Ani Nenkova. 2011a. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 605–613.

Annie Louis and Ani Nenkova. 2011b. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Annie Louis and Ani Nenkova. 2013a. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2013b. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156.

Xiaoyi Ma. 2004. Multiple-Translation Chinese (MTC) Part 3 LDC2004T07. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Xiaoyi Ma. 2006. Multiple-Translation Chinese (MTC) Part 4 LDC2006T04. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Guenter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics - Special issue on using large corpora*, 19(2):313–330.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-08: HLT*, pages 1003–1011.

Thomas A. Mathew and E. Graham Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium*, pages 2–3.

Andrew McKinlay and Katja Markert. 2011. Modelling entity instantiations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 268–274.

Andrew James McKinlay. 2013. *Modeling Entity Instantiations*. Ph.D. thesis, The University of Leeds.

Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 275–286. Springer Berlin Heidelberg.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135.

Kshitij Mishra, Ankush Soni, Rahul Sharma, and Dipti Sharma. 2014. Exploring the effects of sentence simplification on Hindi to English machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society*.

Vivi Nastase, Alex Judea, Katja, and Michael Strube. 2012. Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 183–193.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 896–903.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing*, 12:1532–1543.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference: Short Papers*, pages 13–16.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the International Conference on Computational Linguistics: Companion volume: Posters*, pages 87–90.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691.

Rashmi Prasad and Aravind Joshi. 2008. A discourse-based approach to generating why-questions from texts. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, pages 1–3.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3030–3037.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, pages 448–453.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136.

Vasile Rus, Philip M. McCarthy, Arthur C. Graesser, and Danielle S. McNamara. 2009. Identification of sentence-to-sentence relations using a textual entailer. *Research on Language and Computation*, 7(2):209–229.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808.

Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. *Linguistic Data Consortium, Philadelphia*.

Christopher Scanlan. 2000. *Reporting and writing: Basics for the 21st century*. Harcourt College Publishers.

RJ Senter and EA Smith. 1967. Automated readability index. Technical report, DTIC Document.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842, December.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165:259–298.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156.

Michael Stinson and Octavious A. Tracy. 1983. Specificity of word meaning and use of sentence context by hearing-impaired adults. *Journal of Communication Disorders*, 16(3):163 – 173.

Andreas Stolcke. 2002. SRILM - An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

Paul Taylor and Alan W. Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117.

Radoslava Trnavac and Maite Taboada. 2013. Discourse relations and affective content in the expression of sentiment in texts. In *11th ICGL Conference–Workshop on The semantic field of emotions: Interdisci.*

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Prroceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.*

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–374.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Kees van Deemter and Stanley Peters. 1996. *Semantic ambiguity and underspecification.* CSLI Publications.

Svitlana Volkova and David Yarowsky. 2014. Improving gender prediction of social media users via weighted annotator rationales. In *NIPS 2014 Workshop on Personalization.*

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Shengqin Xu and Peifeng Li. 2013. Recognizing Chinese elementary discourse unit on comma. In *International Conference on Asian Language Processing*, pages 3–6.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers.*

Nianwen Xue, Fei Xia, Fu-dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238, June.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19.

Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Dapeng Yin, Fuji Ren, Peilin Jiang, and Shingo Kuroiwa. 2007. Chinese complex long sentences processing method for Chinese-Japanese machine translation. In *International Conference on Natural Language Processing and Knowledge Engineering*, pages 170–175.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804.