My research interests are in Natural Language Processing (NLP) with focus on computational discourse, i.e., how information is organized into textual units such as sentences and clauses, and the relationship between these units. The driving passion behind my research is to make textual information more accessible, easier to organize and navigate, and to develop techniques that can be readily applied in inter-disciplinary research. Knowledge of discourse is crucial in determining text structure, evaluating the salience and relevance of information, and assessing coherence and reading complexity. I have developed techniques to analyze discourse structure and quantified the impact of discourse organization on text understanding. I am motivated by the challenges of text understanding above the sentence level (e.g., paragraphs and documents), the need for more coherent and intelligible discourse in text generation, as well as the improvement of communication effectiveness among audiences with distinct backgrounds (e.g., speakers of different languages, people with communication difficulties, etc.).

My contributions include novel techniques and models in three aspects of discourse: text specificity [1, 2, 3], content organization [4, 5, 6, 7] and semantic relationships [3, 5, 8, 9], with applications in text simplification [1], automatic summarization [7], machine translation [4, 6] and health informatics [10]. I pursue and advocate the use of linguistic insights and semi-supervised approaches for low-resource learning problems [1, 3, 5] and the understanding of discourse structure in inter-disciplinary applications and large datasets [7, 10].

## Text specificity

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, hence impacting the coherence and the quality of text. 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. Unspecified text also often relies on context to be fully comprehended.

My work contributes to the efficiency and scalability in specificity prediction and the discourse-level annotation of text specificity. In practice, my work enables speedy incorporation of text specificity in sentence simplification [1], argumentation mining [11] and feedback quality in education [12]. I have been working with the Center for Autism Research at the Children’s Hospital of Philadelphia to investigate the connection between text specificity and communication difficulties in autism [10]. In addition, I have been approached by various parties from education, social science and law to investigate text specificity in their tasks.

### Semi-supervised learning for efficiency and scalability.

To allow other tasks to make use of text specificity, it is important to have an efficient system, especially in our era of increasingly large-scale datasets. My work focuses on sentence specificity, which characterizes the level of detail in a sentence. I target three key aspects of improvement from traditional approaches: a) use lightweight text processing; b) use scalable representations for words; c) bypass expensive human annotation. I developed Speciteller [1], a novel, semi-supervised system that relies only on the string surface 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.

### Data annotation for complex discourse effects.

Lack of specificity can be recognized through various discourse factors, for example, referring expressions to entities or events, adjectives with previously established standards, teaser expressions, etc. My work contributes to establishing reliable annotated data to assess specificity in context. I have developed 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 [2]. The schema revealed high-level reading comprehension effects associated with specificity and the use of words, insights that contribute to substantial improvements when predicting certain discourse relations [3].

### Content organization

Language use is often subject to change depending on the application and the context in which communication takes place. My work studies such changes in the discourse organization of human and machine generated text in two common scenarios: conveying the same semantic content across languages and summarizing content in short texts.

**Cross-lingual discourse divergence and translation.** Languages differ in the use of discourse devices that organize information into complex sentences. For example, the Chinese sentence shown below is expressed in two English sentences. In addition, the **conjunction** relation explicitly expressed in English using the connective “and” (underlined) is implicit in Chinese:

```
来
自
美
国
、
日
本
、
新
加
坡
的
外
资
增
加
较
多
，
新
项
目
中
外
资
投
资
比
例
越
来
越
高
，
独
资
企
业
显
著
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.

Without properly handling discourse-related variabilities across languages, I observed that understanding of translated text can be affected even with human translation [4]. It can be especially problematic for machine translation (MT) systems, which are normally trained to translate a sentence in one language into a single sentence in another. I discovered that about 25% of the Chinese sentences in news text should be translated into multiple sentences in English for proper intelligibility [4, 6]. These sentences are not solely distinguishable by length and their existence effectively blurs the boundary of sentence- and discourse-level processing. I developed a novel system to effectively identify these sentences and showed that the identified sentences involve substantially worse MT performance [4]. I further identified other discourse factors, such as the use of discourse connectives, that are different in the two languages [5] and can significantly impact MT quality [6]. This is the first work in this direction for translation from Chinese to English.

**Discourse units in summarization.** Content reorganization also happens during summarization, when the most important information in a document needs to be expressed in much shorter text. Due to information overload, automatic summarization systems have become highly prevalent in both research and industry. While much research focuses on techniques to extract full sentences from a document, manual summaries involve actions such as trimming content, combining sentences and replacing words and phrases. My work explores discourse organization patterns that help systems to produce more human-like summaries. I showed that linguistic units of discourse (i.e., independent clauses) play a characteristic role in human summarization. Judicious use of these units substantially improves the performance of even the most basic summarization systems [7]. This work was nominated for the SIGDIAL best paper award in 2016.

### Semantic relationships

Recognizing discourse relations, i.e., semantic relationships between clauses or sentences such as contrast and cause, has been a critical aspect of discourse processing regardless of the underlying theory of the structure of discourse units (e.g., graph, tree or linear). Motivated by the scarcity of annotated resources for rare relations and other languages, my work makes use of linguistic knowledge, feature representations and semi-supervised methods to significantly improve monolingual and cross-lingual discourse relation recognition.

**Monolingual discourse relation recognition.** One of the most difficult tasks in discourse processing is predicting implicit relations — those not signaled by a cue word (e.g., “because”, “however”) and therefore need to be inferred. For example the **causal** relation between the two sentences below is implicit:

```
They stopped delivering junk mail. Now thousands of mailers go straight into the trash.
```

Although annotated training data is available for implicit discourse relations, examples for many relations remain scarce though they are often important, e.g., the **contrast** relation in sentiment analysis or the **instan-tiation** relation in text specificity. The performance of machine learning systems consequently suffers from the
lack of sufficient training data for these relations. My work tackles the issue by exploiting linguistic knowledge and augmenting feature representations in learning systems. For the *instantiation* relation, I presented a system backed by characteristics such as word usage, word semantic relationships and lexical similarity that substantially improve the recognition of the relation [3]. I also proposed novel feature representations and methods for feature weight augmentation to balance frequent and infrequent relations during training. These techniques showed significant improvement over multiple relations [8, 9].

**Cross-lingual discourse relation identification.** Reliable automatic discourse processing is possible because of the development of large human annotated corpora. However building these resources, especially at the discourse level, is expensive. Hence it is worthwhile to study techniques to transfer existing annotations in English to other languages; such techniques can be especially useful for languages lacking in large corpora creation efforts. I proposed a novel semi-supervised system for explicit discourse relation classification in Chinese using only an automatic discourse parser in English and parallel corpora [5]. The techniques from this work are further generalizable to other languages.

**Future work**

My research interests lie in understanding the discourse structure of text and its impact on text comprehension. In the future I would like to develop text processing techniques that generalize to more domains and languages, to improve natural language generation systems with knowledge in discourse, and to work with other disciplines in discourse understanding that takes into account distinct backgrounds of the audiences.

**Discourse processing in low-resource situations.** Today’s text processing problems are growing rapidly in diversity with different domains and genres of text: formal, informal, commentary, narrative, etc. Unfortunately, resources to train machine learning systems are slow to build and expensive to obtain. This is especially true for discourse annotation which usually considers larger context; for example, the Penn Discourse Treebank, the largest dataset for discourse relations in English newswire, took years to develop and more to refine. Still, this dataset is only for one domain in one language, and as I have mentioned, training examples remain scarce for certain important relations. I believe novel techniques that require less dedicated training data will continue to be highly impactful. Building on my prior work, I will further explore generalizable models via a) transferring information and knowledge in existing resources across tasks and languages; b) leveraging unlabeled data via semi-supervised learning; and c) domain adaptation. I am further interested in developing deep learning techniques for discourse processing that combine the above mentioned three aspects and incorporate linguistic knowledge.

**Natural language generation.** I am interested in advancing the adoption of discourse aspects in text generation systems. My prior work has validated the critical role of discourse units in editor-produced news summaries [7]. I plan to investigate concept extraction and organization patterns in generating coherent summaries. These patterns can involve background knowledge, information salience and text characteristics such as style and specificity. Furthermore, I have a strong interest in making information more accessible, for example, by aiding reading comprehension and language learners. For text simplification, I showed that a decrease in specificity is one of the characteristics of simplified sentences and in determining which sentences need simplification [1]. I also observed that the sentence-discourse discrepancy, similar to that in Chinese-English translation, happens often during simplification. I would like to combine the two lines of work to improve automated text simplification. In addition, I would like to build systems that help non-expert readers by a) rewriting specific text to general while controlling for vagueness and understandability, and b) identifying domain-specific terms and providing relevant details.

**Perspective-dependent discourse and inter-disciplinary research.** The same textual content can be comprehended very differently depending on a person’s background. In my prior work I have showed that the same discourse organization can be well accepted in one language but not in another [4, 6]; the level of specificity should be reduced for text to be more accessible [1]. I plan to take these findings further into aspects in socio-demographics and personal background. Research in social media text has found distinctive language usage across people of different genders, income levels, personality and political views. However, current work focuses mainly on word usage; instead I would like to investigate how discourse is perceived and organized when these aspects vary. My preliminary work with the Center for Autism Research at the Children’s Hospital of Philadelphia investigates text specificity perception and production for people.
with higher traits of autism [10]. I look forward to broader collaboration with researchers from different disciplines such as healthcare, education, digital humanities and social science for perspective-dependent discourse understanding.

References


[7] Junyi Jessy Li, Kapil Thadani, and Amanda Stent. The role of discourse units in near-extractive summarization. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2016 (This paper was nominated for the Best Paper Award).


