Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

# Estimating Text Intelligibility via Information Packaging Analysis

**Junyi Jessy Li**
University of Pennsylvania
Department of Computer and Information Science
Philadelphia, PA 19104
ljunyi@seas.upenn.edu

## Introduction

Effective communication through language involves organizing the content a person or system wishes to convey into text that flows naturally. There are many ways to render the same information, but those appropriate for one group of audience may not be intelligible to another. For example, an English speaker may find it difficult to process a sentence in subject-object-verb order. A child may not understand a complex sentence with multiple clauses. In these situations, it is important for both human and text generation systems to properly reorganize text segments to be well understood. While sentence-level elements such as syntax are important to ensure grammaticality, text comprehension is much determined by how information is arranged into discourse. Consider the following snippets:

> 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*)

To make the text accessible to kids while keeping it informative, the authors of Britannica Elementary selectively removed some information and repacked the rest into two different sentences. The resulting text is thus different in both its *discourse structure* (two sentences vs. one) and its *specificity* (general vs. detailed).

The goal of this thesis to analyze and address factors that influence the intelligibility of text from two aspects of information packaging: discourse structure and text specificity. In particular, I aim to identify discourse phenomena that are highly significant for the quality of text generation system outputs in both cross-lingual and monolingual context. I will also introduce new dataset and methods for analyzing text specificity in its discourse. Using this corpus, I seek to propose techniques that improve the organization of system outputs at multi-sentence level. In addition to automatic text generation systems, my research can also be applied to providing feedback to authors on the naturalness and flow of their text, which currently is difficult to obtain without an editor.

## Discourse Structure Variance

Discourse structure represents the organization of text in fine grained units such as sentences or sub-sentential units (e.g. clauses), and relationships between them (e.g. causal or contrast). Well-packaged discourse ensures a natural flow of the text; unusual or bad structure, on the other hand, renders the text incoherent and unintelligible. In this thesis, I would like to identify factors in discourse structure that are most important when text is reorganized. In doing so, systems can adopt necessary processing to ensure text intelligibility. Writers can also be informed of potential discourse structuring problems in their text.

**Cross-lingual analysis.** Differences in languages involve aspects at various granularity. At the word level, vocabularies, including morphology, are different; at the sentence level, phrases are ordered differently. Discourse structure is no exception (Marcu, Carlson, and Watanabe 2000). Each of these aspects impact machine translation (MT) systems. However, although much prior work focused on the word or sentence level, discourse structure variances were rarely analyzed. This is mainly due to the fact that MT systems standardly translate one input sentence into a single sentence in the target language, forgoing the multi-sentence nature of discourse structure. In this thesis I would like to show that discourse phenomena are highly influential for system translations to be well comprehended.

Across languages, the amount of content that can be reasonably packaged into a single sentence varies. The result is that information originally packaged in a single sentence in one language sometimes must be expressed as a multi-sentence discourse in another. Failure in doing so may result in a translated sentence being hard to process for speakers in the target language. To understand whether this phenomenon is important enough for human and system translators to consider, I showed that sentences need to be translated into multiple English sentences cause significant quality drop for MT, while the number of words in them has little correlation with MT quality (Li, Carpuat, and Nenkova 2014; Li and Nenkova 2015a). From a translator's point of view, more than 15% of the sentences were translated into multiple sentences in English by at least three out of four human translators. From a reader's point of view, for more than 27% of the sentences, at least three out of five readers prefer a multi-sentence translation. Therefore discourse structure should no longer be irrelevant for MT; indeed, we have identified a series of factors associated with the realization of discourse relations that trigger significantly more manual post edits (Li, Carpuat, and Nenkova 2014).

To identify sentences that need a multi-sentence translation, I designed a system that achieves more than 80% accuracy (Li and Nenkova 2015a). Next I plan to explore methods to improve the flow of MT outputs for these sentences. For example, sentence splitting can be adopted prior to translation. Alternatively, the outputs can be post-processed for the edits needed.

**Monolingual analysis.** The necessity of repackaging the content of one sentence into multiple ones in fact reflects that the original sentence does not follow the text flow the target audience group is accustomed to. Therefore it is not exclusive in cross-lingual context: as illustrated in the example in the Introduction section, a complex sentence often needs to be decomposed into multiple sentences for it to be intelligible to kids. For a writer, it is also a specific indication that the writing quality can be improved if the writer rearranges the sentence. This can be particularly helpful for non-native speakers writing in English, whose native language may involve discourse structure very different from that in English. I will generalize my study of sentence-discourse discrepancy to text simplification and discourse organization quality estimation in English.

## General and Specific Content in Text

Text with high intelligibility must flow naturally as well as effectively communicate the content within. In many situations, it is necessary to organize content into a new piece of text with different amount of detail, or *specificity*. In multi-document summarization, human-written abstract summaries are shown to be much less detailed than their machine-generated counterparts which extract raw sentences from the input (Louis and Nenkova 2011). In text simplification, I found that simplified sentences are significantly less specific (Li and Nenkova 2015b). To properly convey the information in the entire document, it is insufficient to arrange text segments using its specificity level without consideration of discourse. Thus I will utilize contextual information in text specificity prediction and analysis.

**A corpus for context-informed sentence specificity.** In my prior work I have improved sentence specificity prediction: one specificity score is assigned to each sentence (Li and Nenkova 2015b). However, a per-sentence model does not consider the specificity status of text units within a sentence and its relationship with respect to the running discourse. Therefore it cannot help us to understand the importance and purpose of information expressed in a sentence.

Currently, we are compiling a corpus of sentence specificity annotation in the context of its original article. The goal of this corpus is to establish a link between the specificity status of sub-sentential expressions and multi-sentential discourse. We have designed an annotation process to identify the *scope* and *aspect* of each underspecified parts in the sentence. In particular, the *scope* records whether the information needed to clearify the underspecification exists in the document context (ie. is context-dependent). The *aspect* records the type of information that is missing. We have collected annotations of 543 sentences (15K words) from 16 politics and business news articles. Each sentence is annotated by three native English speakers and in total they identified and annotated nearly 3K underspecified parts.

**Sub-sentential text units and information packaging.** With the availability of the above annotation, text specificity can now be analyzed at sub-sentential level. The corpus also allows us to factorize the content of text in terms of its former mentions in an article. For underspecified parts that are context-dependent, I will study their relationship with traditional anaphora resolution. For those that are context independent, I will inspect if they represent new or generic information that anchors the planning of a sentence or higher level discourse. I plan to propose methods that identify fully specified and underspecified parts in a sentence and disambiguate the two types of underspecification.

The salience of information present in a sub-sentential unit can now not only be described in its discourse structure, but also in its status of specificity. I plan to develop systems that identify parts of text that most effectively serve the communication purpose at hand, therefore guiding discourse restructure processes to produce output of higher intelligibility. For automatic summarization systems, sub-sentential specificity can aid in deciding whether to include curtain information or not. For sentence simplification systems, deleting details of unimportance and specifying parts difficult to understand are equally important. Previously I found that specificity is indicative in deciding whether a sentence needs to be simplified. I now plan to study the role of sub-sentential specificity when decomposing content into multiple sentences.

## Conclusion: Current and Future Work

My research focuses on analyzing information packaging factors for text intelligibility. I have shown that discourse structure is crucial for well-versed translation and improved the prediction of sentence specificity. We are composing a context-informed text specificity corpus. My future work includes analyzing discourse structure for text intelligibility estimation in English, sub-sentential specificity analysis and its implication for discourse restructuring.

## References

Li, J. J., and Nenkova, A. 2015a. Detecting content-heavy sentences: A cross-language case study. In *Proceedings of EMNLP*.

Li, J. J., and Nenkova, A. 2015b. Fast and accurate prediction of sentence specificity. In *Proceedings of AAAI*.

Li, J. J.; Carpuat, M.; and Nenkova, A. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of ACL, Short Papers*.

Louis, A., and Nenkova, A. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation (MTTG)*.

Marcu, D.; Carlson, L.; and Watanabe, M. 2000. The automatic translation of discourse structures. In *Proceedings of NAACL*.