Towards Domain-Independent Text Structuring Trainable on Large Discourse Treebanks

Grigorii Guz and Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
{gguz, carenini}@cs.ubc.ca

Abstract

Text structuring is a fundamental step in NLG, especially when generating multi-sentential text. With the goal of fostering more general and data-driven approaches to text structuring, we propose the new and domain-independent NLG task of structuring and ordering a (possibly large) set of EDUs. We then present a solution for this task that combines neural dependency tree induction with pointer networks and can be trained on large discourse treebanks that have only recently become available. Further, we propose a new evaluation metric that is arguably more suitable for our new task compared to existing content ordering metrics. Finally, we empirically show that our approach outperforms competitive alternatives on the proposed measure and is equivalent in performance with respect to previously established measures.

1 Introduction

Natural Language Generation (NLG) plays a fundamental role in data-to-text tasks like automatically producing soccer, weather and financial reports (Chen and Mooney, 2008; Plachouras et al., 2016; Balakrishnan et al., 2019), as well as in text-to-text generation tasks like summarization (Nenkova and McKeown, 2012).

Generally speaking, NLG involves three key steps (Gatt and Krahmer, 2017): first there is content determination which selects what information units should be conveyed, secondly there is text structuring, which is responsible for properly structuring and ordering those units; and finally microplanning-realization that aggregates information units into sentences and paragraphs that are then verbalized.

The focus of this paper is on the text structuring step, which is critical for the overall performance of an NLG system, as it ensures that the communicative goals of the text are realized in the most structurally coherent and cohesive way possible, making the main ideas expressed by the text easy to follow for the target audience.

Aiming to develop very general computational methods for text structuring, we keep our study independent from particular ways in which the input information units are represented and from explicitly provided ordering constraints for the target application domain (Gatt and Krahmer, 2017). More specifically, we propose and attack, in a fully data-driven way, the novel and domain-independent task of simultaneously structuring and ordering a set of Elementary Discourse Units (EDUs), i.e., clause-like text fragments that the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) assumes to be the building blocks of any discourse structure (see Figure 1(a)(left)). In other words, we assume that the system is given a set of EDUs (with cardinality possibly > 100) as input and returns their ordering, as well as the unlabelled RST dependency discourse tree structure for a document consisting of this set of EDUs, as illustrated in Figure 1(a).

Our data-driven approach relies on the very recent availability of large treebanks containing hundreds of thousands of (silver-standard) discourse trees that can be automatically generated by distant supervision following the approach presented by Huber and Carenini (2020). We formulate the problem as one of the dependency tree induction, repurposing existing solutions (Ma and Hovy, 2017; Vinyals et al., 2015) to perform an RST-based text structuring where both EDU ordering and tree building are executed simultaneously (Reiter and Dale, 2000). The resulting structures can be highly useful for subsequent NLG pipeline stages such as aggregation, and for downstream tasks like text simplification (Zhong et al., 2019). Our approach is trainable end-to-end, but since the discourse trees in the training treebank are constituency trees (see

Figure 1(b)), we face the additional challenge of turning them into dependency trees (see Figure 1(a)) before the learning process can start (Hayashi et al., 2016).

In a comprehensive evaluation, we compare our solution to three baselines along with a competitive approach based on pointer networks (Vinyals et al., 2015), which is the established method of choice not only for sentence ordering (Cui et al., 2018), but also for basic domain-specific text structuring in data-to-text applications (Puduppully et al., 2019). In particular, the comparison involves training and testing the different models on the MEGA-DT treebank (Huber and Carenini, 2020), containing ≈250,000 discourse trees obtained by distant supervision from a the Yelp'13 corpus of customer reviews (Tang et al., 2015).

With respect to evaluation metrics, we found the current ways of measuring content ordering (e.g., Kendall's τ) to be inadequate to capture the quality of long sequences of relatively short information units (i.e., sequences of EDUs of long multi-sentential text). Thus, we propose a novel evaluation measure, Blocked Kendall's τ, that we argue should be used for our new NLG task of ordering and structuring a possibly large set of EDUs, because it critically measures how well semantically close units are clustered together in the correct order.

To summarize the contributions of this paper: **(i)** we propose the new and domain-independent NLG task involving the structuring and ordering a set of EDUs, which is intended to enable more general and data-driven approaches to text structuring; **(ii)** we present a strong benchmark solution for this task, trainable on large discourse treebank, that combines neural dependency tree induction with pointer networks; **(iii)** we propose a new evaluation metric that is arguably much more suitable for this task than existing ordering metrics; **(iv)** and on this new metric along with standard tree-quality metrics, we show empirically that our approach outperforms or is comparable to competitive alternatives. The code for our solution and the new metric, as well as the treebank for training, is publicly available.[1]

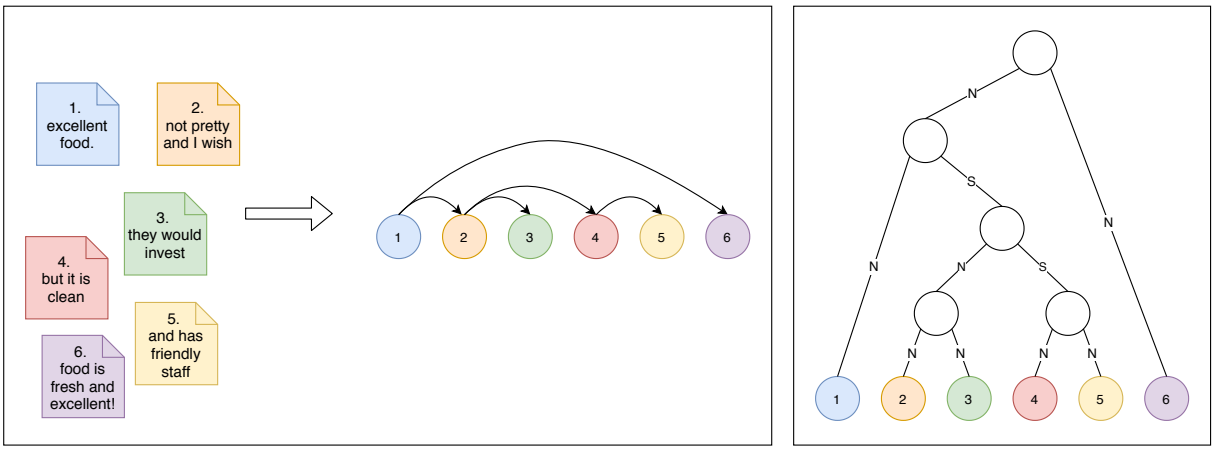
[1] http://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/index.html

2 Related Work

(a) Text structuring is a key step in NLG, especially when generating long multi-sentential documents. Not surprisingly, this is also the case in recent neural approaches. Wiseman et al. (2017) presented the RotoWire corpus, targeting long-document data-to-text NLG. To generate the document, their model conditions on all records in the data table by weighting their embeddings with attention, in addition to using copying mechanism for out-of-vocabulary data entries. The follow-up work of Puduppully et al. (2019), instead of conditioning on all records, arguably performs better text structuring by first selecting and then ordering the entries of a data table using Pointer network architecture (Vinyals et al., 2015). That way, the surface realization module considers previously generated text and only one new data table entry at a time. Their model was extended by Iso et al. (2019), with an additional GRU for tracking the entities that the model already referred to in the past. Pursuing a rather different approach to improve text structuring, Shao et al. (2019) proposed a hierarchical latent-variable model where the problem is decomposed into dependent sub-tasks, aggregating groups of data table entries into sentences first and then generating the sentences sequentially, conditioned on the plan and already generated sentences. Overall, these last three models significantly outperform the initial approach of Wiseman et al. (2017) both in terms of fluency and coverage, with increasing sophistication of the text structuring module yielding bigger gains, confirming that text structuring is indeed crucial for generating coherent long documents.

The task we propose and investigate in this paper can be seen as pushing this line of research even further. We aim for a more ambitious text structuring module inspired by traditional NLG work, viewing the process as the construction of an RST discourse tree for the target document (Reiter and Dale, 2000), which critically includes assigning importance to each constituent. Tellingly, our task is also domain-independent and agnostic on the representation of the input information units.

(b) The goal of **sentence ordering** is to sort a given set of unordered sentences into a maximally coherent document. Most recent work on sentence ordering (Logeswaran et al., 2016; Cui et al., 2018; Wang and Wan, 2019) involves constructing contextualized order-agnostic representations of indi-



(a) (b)

Figure 1: (a) A simple example of the novel NLG task we propose in this paper: generating an ordered discourse dependency tree (right) for a given set of EDUs (left). (b) The constituency discourse tree corresponding to the dependency tree shown in (a). The RST-style discourse trees in the treebanks we use for our experiments are initially represented as constituency trees.

vidual sentences and full documents using architectures such as Transformer Encoder without positional embeddings (Vaswani et al., 2017), and then feeding those representations into a pointer-based decoder (Vinyals et al., 2015).

The new task we propose in this paper is similar, but more challenging than sentence ordering. Instead of ordering sentences, we need to order EDUs, which are often shorter sentence constituents, and therefore by expressing smaller semantic units they arguably require more fine-grained processing. Furthermore, our task goes beyond ordering by also requiring the synergistic and simultaneous step of generating the RST discourse structure for the EDUs. To address these challenges, more powerful techniques for tree induction are needed on top of pointer networks.

(c) Document discourse tree structure induction: The third related line of research involves the induction of latent tree structures over documents. Some of these works aim at obtaining better document representations for tasks such as text classification (Karimi and Tang, 2019) and single-document extractive summarization (Liu et al., 2019). In essence, a neural framework is designed so that a discourse tree for a document is induced while training on the target downstream task. However, even if these approaches demonstrated improvements over non-tree-based models, subsequent studies have shown that the resulting latent discourse dependency trees are often trivial and too shallow (Ferracane et al., 2019). In contrast, recent work on distant supervision from sentiment (Huber and Carenini, 2020) indicates that large treebanks of discourse trees can be generated by combining neural multiple-instance learning (Angelidis and Lapata, 2018) with a CKY-inspired algorithm (Jurafsky and Martin, 2014). Since a series of experiments in inter-domain discourse parsing have certified the high-quality of these treebanks, we use one of such treebaks, called MEGA-DT, for training and testing our data-driven text structuring approach.

3 Novel Task and Methods

Our novel task for text structuring receives as input a set of n EDUs and returns both an ordering and a discourse structure for that set. We first describe how the EDUs are encoded, as this is the initial step for all the approaches we consider. Then, after discussing a basic method for just ordering the input EDUs (Pointer Networks), which will serve as our main baseline, we present our solution for fully solving the task in detail, which combines tree induction with pointer networks. We will refer to our final approach as *DepStructurer*. We conclude the section with two simple baselines for EDU ordering and structuring, respectively.

3.1 EDU Encoder

For a clear comparison of tree vs. non-tree based approaches, we encode EDUs in a very similar way to previous sentence ordering works (Cui et al., 2018; Wang and Wan, 2019). Given a document