Discourse Planning with an N-gram Model of Relations

Or Biran
Columbia University
orb@cs.columbia.edu

Kathleen McKeown
Columbia University
kathy@cs.columbia.edu

Abstract

While it has been established that transitions between discourse relations are important for coherence, such information has not so far been used to aid in language generation. We introduce an approach to discourse planning for concept-to-text generation systems which simultaneously determines the order of messages and the discourse relations between them. This approach makes it straightforward to use statistical transition models, such as n-gram models of discourse relations learned from an annotated corpus. We show that using such a model significantly improves the quality of the generated text as judged by humans.

1 Introduction

Discourse planning is a subtask of Natural Language Generation (NLG), concerned with determining the ordering of messages in a document and the discourse relations that hold among them (Reiter and Dale, 2000). Early approaches to discourse planning used manually written rules, often based on schemas (McKeown, 1985) or on Rhetorical Structure Theory (RST) (Mann and Thompson, 1987; Hovy, 1993; Power, 2000). In the past decade, various statistical approaches have emerged (Duboue and McKeown, 2001; Dimitromanolaki and Androutsopoulos, 2003; Soricut and Marcu, 2006; Konstas and Lapata, 2013). Other relevant statistical approaches to content ordering can also be found in the summarization literature (Barzilay et al., 2001; Lapata, 2003; Bollegala et al., 2005). These approaches overwhelmingly focus on determining the best order of messages using semantic conent, while discourse relations are in most cases either determined by manually-written derivation rules or completely ignored.

Meanwhile, researchers working on discourse relation disambiguation have observed that the sequence of discourse relations itself, independently of content, helps in disambiguating adjacent relations (Wellner et al., 2006; Pitler et al., 2008). Sequential discourse information has been used successfully in discourse parsing (Ghosh et al., 2011; Feng and Hirst, 2014), and discourse structure was shown to be as important for text coherence as entity-based content structure (Lin et al., 2011; Feng et al., 2014). Surprisingly, so far, discourse sequential information from existing discourse-annotated corpora, such as the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) has not been used in generation.

In this paper, we present an NLG framework that generates texts from existing semantic web ontologies. We use an n-gram model of discourse relations to perform discourse planning for these stories. Through a crowd-sourced human evaluation, we show that the ordering of our documents and the choice of discourse relations is significantly better when using this model.

2 Generation Framework

In concept-to-text generation pipelines, discourse planning typically occurs after the content selection stage. The input, therefore, is an unordered set of messages that are not yet realized: instead of being represented as text, the messages have a structured semantic representation.

In this paper, we generate *comparison stories*, describing and comparing two similar entities, from an RDF ontology. The RDF semantic representation is commonly used in semantic web resources and free ontologies. An RDF message (called a *triple*) has three parts: a subject, a predicate and an object. For each story, we consider any triple whose subject is one of the participating entities as a potential message to be generated. We do only minimal processing on these messages:

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1973–1977,
Lisbon, Portugal, 17-21 September 2015.

where two triples have the same subject and predicate but different objects, we merge them into a single message with multiple objects; and where two triples have the same subject and object but different predicates, we merge them into a single message with multiple predicates.

Next, we build the set of potential discourse relations between all messages. We use the PDTB *class-level* relations, of which there are four: *expansion*, *comparison*, *contingency* and *temporal*. Each is an abstraction of a family of more specific relations, such as *cause*, *concession*, etc. We do not differentiate between explicit and implicit relations, and treat *entrel* as a type of *expansion*.

Potential discourse relations are implied in the semantics of the triples: messages that contain the same predicate and object may have an expansion relation among them (e.g. "John has a ball. Mary also has a ball"). Messages that contain the same predicate but different subjects and objects may have a comparison relation (e.g. "John likes apples but Mary likes oranges").

Specific predicate pairs will also have specific potential relations among them - for example, "birth place" and "residence" have a temporal relation (when applied to the same subject). The same is true for contingency relations (e.g., "city" and "country" for the same subject - if the subject is in a city, it implies which country it is in). We manually annotated the 59 predicate pairs that had potential temporal and contingency relations, as well as 8 pairs with special potential comparison relations (e.g., "birth place" and "residence" if the subject is the same but the object is not).

Once the potential relations are identified, we have a directed multigraph where each vertex is a message and each edge is a potential relation. There can be multiple edges between any two vertices, since messages may have more than one potential relation among them.

Once the graph is ready, we perform content selection. Given a desired number of messages to generate, we choose the set of messages that maximizes the number of edges in the resulting subgraph (thus ensuring that the selected messages are discourse-coherent). If there are multiple such sets, we choose one at random.

The task we are focused on in this paper is discourse planning, which in our formulation is the task of finding a Hamiltonian path through the selected subgraph, thus simultaneously selecting the order of the messages (nodes) as well as the relations (edges) that connect them. Our approach for choosing the best path is discussed in the next section. For the remainder of this section, we describe our simple implementations of the next stages of generation: sentence planning and realization.

For each of the four discourse relations we use, we selected a few explicit connectives from the PDTB that are often used to convey them. We specifically chose connectives that apply to the entire range of class-level relations (e.g., for *comparison* we chose "while" - since it applies to both *contrast* and *concession* in the PDTB, but not "in contrast" which applies only to the former). We also chose only those connectives which have the structure [ARG1 connective ARG2] or [ARG1. connective, ARG2]. During realization, we arbitrarily choose a connective to realize the relation.

Since the ordering and relations between messages is determined by the discourse plan, sentence planning falls naturally out of it: sentence breaks occur where the connective pattern creates them, or where there is no relation between adjacent messages.

To realize the messages themselves, we follow a single pattern: "the [predicate(s)] of [subject] (is/are) [object(s)]". Simple rules are used to pluralize the predicate when there are multiple objects and to create lists of multiples objects or predicates where needed.

These basic solutions for the various stages of NLG produce texts that are rich enough to be acceptable for human readers, but which have relatively little variation in grammatical and lexical quality. This crucial combination allows us to perform a human study to specifically evaluate the discourse planning component.

3 Discourse Planning

As explained in the previous section, we formulate the discourse planning task as finding a path through a multigraph of potential relations between messages. One major component of what makes a good path is the sequence of content: some content is more central and should appear earlier, for example; and some predicates and objects are semantically related and should appear near one another. In this paper we focus on a component that has so far been neglected in generation - the sequence of discourse relations - while trying to minimize the effect that content semantics

have on the evaluation (other than the semantics implicit in the relations). In order to quantify the likelihood of a sequence of relations, we build an n-gram model from a discourse-annotated corpus.

An n-gram model measures the transitional probabilities for sequences of the units that the n-grams are composed of. In this case, the units are discourse relations. The probability of a particular sequence of relations of length $n+1$ given an existing subsequence of length n is computed as a fraction of the number of times it appears in the corpus and the number of times the subsequence appears in the corpus, i.e.

$$P(r_i|r_{i-n},...,r_{i-1}) = \frac{C(r_{i-n},...,r_{i-1},r_i)}{C(r_{i-n},...,r_{i-1})}$$

Where $C(s)$ is the number of times sequence s appears in the corpus. Using this model to generate a discourse plan given a potential relation multigraph is a stochastic process: at each stage, we choose the next relation edge out of the last chosen message vertex (the first vertex is chosen at random) based on the selected sequence of relation edges and the probabilities for the next relation in the model. Once a vertex is added to the path edges leading to it can no longer be selected.

4 Evaluation

One method for evaluating a discourse plan independently of content is to produce pairs of generated short text documents, each containing the same content, but with different ordering and discourse relations (as dictated by the discourse plan). The only obvious way to decide which text is better is to have human judges make that decision. It is important to minimize the effects of other qualities of the texts (differences in content, word choice, grammatical style, etc.) as much as possible, so that the judgement is based only on the differences in order and discourse.

We used DBPedia (Auer et al., 2007) - an RDF ontology extracted from Wikipedia - to generate content. Each document generated was a *comparison story* of two entities in a single category (i.e., the messages in the stories were selected, as described in Section 2, from the set of triples where one of the entities was the subject). In order to experiment with different domains, we used four different categories: **Office Holder** (i.e., a person holding office such as a President or a Judge); **River**; **Television Show**; and **Military Unit**. The entity pairs from each category were chosen at random but were required to have at least 8 predicates and 3 objects in common, so that they were somewhat semantically related.

The birth place of Allen J. Ellender is Montegut, Louisiana, while the death place of Allen J. Ellender is Maryland. The birth place of Robert E. Quinn is Phoenix, Rhode Island. Subsequently, the death place of Robert E. Quinn is Rhode Island.
The birth place of Allen J. Ellender is Montegut, Louisiana. In comparison, the birth place of Robert E. Quinn is Phoenix, Rhode Island. The death place of Robert E. Quinn is Rhode Island, but the death place of Allen J. Ellender is Maryland.

Figure 1: Sample pair of comparison stories

To ensure that human judges can easily tell the differences between the stories on a sentential level, we limited the size of each story to 4 messages. For each pair of stories, everything but the discourse plan (i.e. the content selection, the realization of messages and the lexical choice of connectives) was identical. Figure 1 shows an example pair of stories from the Office Holder category.

4.1 Experiments

We conducted two crowd sourced experiments on the CrowdFlower platform. Each question consisted of two short stories that are completely identical in content, but each generated with a different discourse planner. The human judge was asked to decide which of the stories has a better flow (or whether they are equally good), and then to give each of the stories a score from 1 to 5, paying specific attention to the ordering of the prepositions and the relations between them. The stories were presented in a random order and were not given labels, to avoid bias. We generated 125 pairs of stories from each category - a total of 500 - for each experiment.

Each question was presented to three judges. In each experiment, there was complete disagreement among the three annotators in approximately 15% of the questions, and those were discarded. In approximately 20% there was complete agreement, and in the rest of the questions there were two judges who agreed and one who disagreed. We also computed inter-annotator agreement using Cohen's Kappa for 217 pairs of judges who

	Quality comparison			Avg. score	
	Base	Equal	Pdtb	Base	Pdtb
Of. Holder	27.4%	30.2%	**42.5%**	3.67	**3.76**
TV Show	34.3%	25.7%	**40%**	3.79	**3.8**
Mil. Unit	32.3%	23.2%	**44.4%**	3.69	**3.84**
River	**39.2%**	23.5%	37.3%	3.71	**3.72**
Total	34%	25%	**41%**	3.72	**3.78**

Table 1: Results for the comparison between the PDTB n-gram model and the baseline

	Quality comparison			Avg. score	
	Pdtb	Equal	Wiki	Pdtb	Wiki
Of. Holder	33.6%	14.5%	**51.8%**	3.51	**3.65**
TV Show	43.2%	8.1%	**48.6%**	3.62	**3.65**
Mil. Unit	40.4%	14.4%	**45.2%**	3.65	**3.67**
River	41.1%	11.2%	**47.7%**	3.68	**3.7**
Total	39.6%	12%	**48.4%**	3.61	**3.67**

Table 2: Results for the comparison between the Wikipedia model and the PDTB model

both answered at least 10 of the same questions. The average kappa value was 0.5, suggesting reasonable agreement.

In the first experiment, we compared stories generated by a planner using an n-gram model extracted from the PDTB with stories generated by a baseline planner, where all edges have identical probabilities. The results are shown in Table 1.

In the second experiment, we used a PDTB shallow discourse parser we developed (Biran and McKeown, 2015) to create a discourse-annotated version of the English Wikipedia. We then compared stories generated by a planner using an n-gram model extracted from the parsed Wikipedia corpus with those generated by a planner using the PDTB model. The results are shown in Table 2.

The total results in both tables are statistically significant ($p < 0.05$).

4.2 Discussion

The results in Table 1 show that the judges significantly preferred the stories created by the n-gram model-based planner to those created by the baseline planner, both in terms of the three-way decision and in terms of the numeric score. This is true for the total set as well as every specific topic, except for *River*. This may be because the predicates in the River category are much more cohesive than in other categories: virtually all predicates related to rivers describe an aspect of the location of the river. That fact may make it easier for a random planner to produce a story that seems coherent. Note, however, that while the judges preferred the baseline story more often in the *River* questions, the average score is higher for the model, which suggests that when the baseline was better it was only mildly so, while when the model was better is was significantly so.

The results in Table 2 show that the Wikipedia-based model produces better results than the PDTB-based model. We hypothesize that it is for two reasons. First, Wikipedia contains definitional texts and is closer in style and content to the stories we produce than the PDTB, which contains WSJ articles. Temporal relations constitute about 10% of both corpora, but contingency and comparison relations each make up almost 20% of the PDTB, while in Wikipedia they span only 10% and 12% of the corpus, respectively, making the share of expansion relations much larger. Second, since the PDTB is small, higher-order n-grams are sparsely found, which can add noise to the model. The Wikipedia corpus is significantly larger and does not suffer from this problem.

The differences in average scores seen in the experiments are relatively small. That is expected, since we have eliminated the content coherence factor, which is known to be significant. In addition, while judges were specifically asked to focus on the order of messages and relations between them, there is inevitably some noise due to accidental lexical or syntactic mismatches, ordering that is awkward content-wise, and other side-effects of the generation framework we employed.

5 Conclusion and Future Work

We introduced an approach to discourse planning that relies on a potential discourse multigraph, allowing for an n-gram model of relations to drive the discourse plan and efficiently determine both the ordering and the relations between messages.

We conducted two experiments, comparing stories generated with different discourse planners. The first shows that an n-gram model-based planner significantly outperforms the random baseline. The second suggests that using an n-gram model derived from a corpus that is larger and closer in style and content, though less accurately annotated, can further improve results.

In future work, we intend to combine this discourse-based view of coherence with a content-based view to create a unified statistical discourse planner. In addition, we will explore additional stochastic models of discourse that look at other, non-sequential collocational information.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2015, Prague, Czech Republic.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2005. A machine learning approach to sentence ordering for multidocument summarization and its evaluation. In *Natural Language Processing–IJCNLP 2005*, pages 624–635. Springer.

Aggeliki Dimitromanolaki and Ion Androutsopoulos. 2003. Learning to order facts for discourse planning in natural language generation. *arXiv preprint cs/0306062*.

Pablo A. Duboue and Kathleen R. McKeown. 2001. Empirically estimating order constraints for content planning in generation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 172–179, Toulouse, France, July. Association for Computational Linguistics.

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, Baltimore, Maryland, June. Association for Computational Linguistics.

Vanessa Wei Feng, Ziheng Lin, Graeme Hirst, and Singapore Press Holdings. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of the 25th International Conference on Computational Linguistics*.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1071–1079.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artif. Intell.*, 63(1-2):341–385, October.

Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1503–1514, Seattle, Washington, USA, October. Association for Computational Linguistics.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 545–552, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. Technical Report ISI/RS-87-190, ISI.

Kathleen R. McKeown. 1985. Discourse strategies for generating natural-language text. *Artif. Intell.*, 27(1):1–41, September.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations.

Richard Power. 2000. Planning texts by constraint satisfaction. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 642–648, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*, volume 33. Cambridge university press.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 803–810, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Saurí. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 117–125, Stroudsburg, PA, USA. Association for Computational Linguistics.