DISCOFUSE: A Large-Scale Dataset for Discourse-Based Sentence Fusion

Mor Geva *
Tel Aviv University
morgeva@mail.tau.ac.il

Eric Malmi
Google AI
emalmi@google.com

Idan Szpektor
Google AI
szpektor@google.com

Jonathan Berant †
Tel Aviv University,
Allen Institute for AI
joberant@cs.tau.ac.il

Abstract

Sentence fusion is the task of joining several independent sentences into a single coherent text. Current datasets for sentence fusion are small and insufficient for training modern neural models. In this paper, we propose a method for automatically-generating fusion examples from raw text and present DISCOFUSE, a large scale dataset for discourse-based sentence fusion. We author a set of rules for identifying a diverse set of discourse phenomena in raw text, and decomposing the text into two independent sentences. We apply our approach on two document collections: Wikipedia and Sports articles, yielding 60 million fusion examples annotated with discourse information required to reconstruct the fused text. We develop a sequence-to-sequence model on DISCOFUSE and thoroughly analyze its strengths and weaknesses with respect to the various discourse phenomena, using both automatic as well as human evaluation. Finally, we conduct transfer learning experiments with WEB-SPLIT, a recent dataset for text simplification. We show that pretraining on DISCOFUSE substantially improves performance on WEB-SPLIT when viewed as a sentence fusion task.

1 Introduction

Sentence fusion is the task of combining several independent sentences into a single coherent text (Barzilay and McKeown, 2005). Sentence fusion is important in many NLP applications, including retrieval-based dialogue (Song et al., 2018; Yan and Zhao, 2018), text summarization (Barzilay and McKeown, 2005; Bing et al., 2015) and question answering (Li et al., 2018; Marsi and Krahmer, 2005). Such systems retrieve multiple sentences from different sources, documents or paragraphs, and use them to construct a coherent text.

> As a run-blocker, **Zeitler** moves relatively well.
>
> **Zeitler** too often struggles at the point of contact in space.
>
> As a run-blocker, **Zeitler** moves relatively well. **However, he** too often struggles at the point of contact in space.
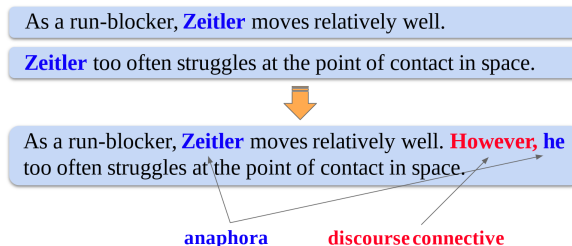>
> **anaphora**   **discourse connective**

Figure 1: Example for two independent sentences, and their fusion. The modifications applied are pronominalization (blue) and connective insertion (red).

Sentence fusion is challenging because it requires understanding the discourse semantics between the input sentences. Consider the example in Figure 1: a coherent fusion of the sentences requires understanding that the second sentence *contrasts* the first one, in order to insert the discourse connective *"However"*. In addition, the gender and syntactic role of the entity *"Zeitler"* needs to be inferred to insert the pronoun *"he"*.

Prior work on sentence fusion (Barzilay and McKeown, 2005; Turner and Charniak, 2005; Filippova, 2010; Elsner and Santhanam, 2011; Thadani and McKeown, 2013; Bing et al., 2015; Chali et al., 2017) utilized very small amounts of labeled data, which are insufficient to train modern neural models. In this work, we propose a method for automatically generating sentence fusion examples at scale from raw text corpora.

To this end, we go over sentences and contiguous pairs of sentences in a corpus, and apply a set of manually-constructed rules, which identify the occurrence of prevalent fusion operations. The rules specify how to modify the sentences such that they are "unfused" into two independent sentences. E.g., in Figure 1 one rule will delete the discourse connective *"However"*, and another will replace the pronoun *"he"* with the named entity *"Zeitler"*.

In the generated examples, the original fused text becomes the target, and the unfused sentences (generated by rules) are the input. Importantly, sentence fusion models trained on our data cannot simply learn to invert rule application, because information is lost and can be recovered only by understanding the text semantics . As mentioned, learning to insert *"However"* in Figure 1 requires inferring that the sentences contrast. We cover a wide range of fusion phenomena such as inserting discourse connectives in various positions of the sentences, anaphora and cataphora identification, and sentence merging through coordination, relative clauses and apposition.

We applied our method on two large document collections, Wikipedia and sports articles from the Web, resulting in two datasets of 16 million and 44 million examples respectively. We call the combined dataset DISCOFUSE. We extensively analyze the quality of our dataset with crowdsourcing, and find that workers understand the text after splitting in 85% of the cases, and the other 15% are due to either the original text being unclear or errors in rule application.

We trained a state-of-the-art sequence-to-sequence model (Vaswani et al., 2017) and analyzed the fusion phenomena in which the model struggles. We found that the model succeeds in fusing sentences through structural constructions such as apposition or relative clauses, but performs badly when fusion involves inserting a particular discourse connective, or selecting pronominals.

Last, we performed transfer learning by training on DISCOFUSE and then fine-tuning on a smaller dataset from a different distribution. To this end, we utilize WEBSPLIT, a recent dataset for sentence splitting (Narayan et al., 2017; Aharoni and Goldberg, 2018), viewing WEBSPLIT as a sentence fusion task. We found that pre-training on DISCOFUSE substantially improves the performance of a fusion model in this setup.

To conclude, our contributions are:

1. DISCOFUSE: a dataset of 60 million sentence fusion examples from two different corpora.
2. A method for automatically generating sentence fusion examples from raw text.
3. Automatic and human evaluation of the Transformer model on the fusion task.
4. A transfer learning setting in which model performance improves when pre-trained with DISCOFUSE.

The DISCOFUSE dataset is publicly available at: `https://discofuse.page.link/data`.

2 Background

Existing fusion datasets are small, which is perhaps why only few works have explored the application of supervised models to sentence fusion (Elsner and Santhanam, 2011; Thadani and McKeown, 2013). McKeown et al. (2010) introduced a human-generated corpus of 3,000 examples. Elsner and Santhanam (2011) extracted around 300 fusion examples from pre- and post-editing news articles. Thadani and McKeown (2013) constructed 1,858 examples from summarization tasks. Such datasets are too small to train modern data-hungry neural models.

Related to sentence fusion is its "inverse" task of sentence splitting. Collados (2013) automatically constructed a Spanish simplification dataset by splitting single sentences into several simpler ones. Recently, two larger datasets for text splitting were released (Botha et al., 2018; Narayan et al., 2017; Aharoni and Goldberg, 2018). However, using these datasets for the "mirror" task of sentence fusion is problematic. First, sentence splitting often involves removing content from the original sentence for simplification, and this content is impossible to recover in the fusion direction. Second, these datasets do not focus on discourse and thus prominent discourse phenomena may be missed. Last, our new dataset is more than an order of magnitude larger than the above sentence splitting datasets.

Another related line of recent work focused on predicting discourse connectives between sentences and automatically generating examples from raw text (Liu et al., 2016; Malmi et al., 2018). We substantially expand over those works by handling more diverse linguistic phenomena, such as connectives in single sentences, generating anaphora and cataphora constructions, relative clauses, coordination and more, which are all represented in a single dataset. Moreover, our dataset is 20x larger compared to prior work, allowing us to examine in depth long-tail scenarios.

3 The DISCOFUSE Dataset

We next describe our process for building DISCOFUSE, which contains 60 million sentence fusion examples from two different document collections: Wikipedia and Web articles about sports.
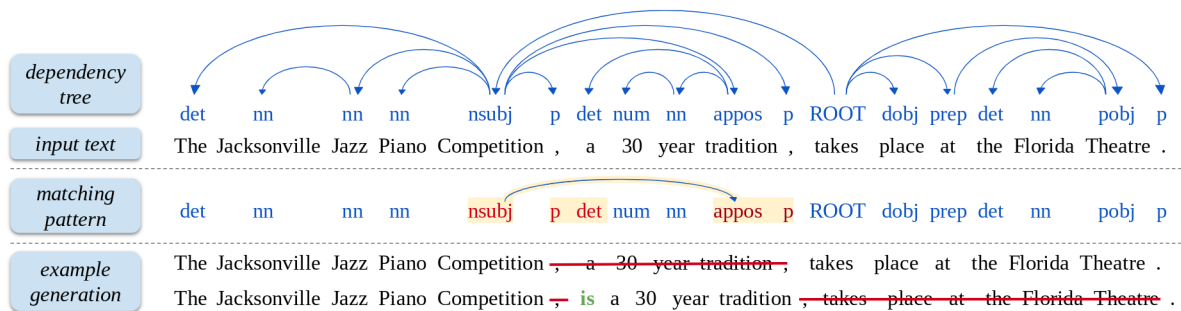
Figure 2: Example generation rule for apposition. Given an input text and its dependency tree, we check for a match with the apposition pattern. We then use the dependency tree to split the sentence and create a new example.

3.1 Example Generation

DISCOFUSE contains union-fusion examples, *i.e.* fusing sentences without loss of content (Marsi and Krahmer, 2005). To automatically extract examples, we manually crafted a list of text splitting rules. Our rule-set covers 9 fusion phenomena, including handling discourse connectives, coordination and relative clauses, and entity resolution for anaphora and cataphora constructions. For entity resolution, both anaphoric pronouns (*"she"*, *"they"*, *"his"*) and anaphoric nominals (*"the team"*, *"the man"*) are considered, based on the output of a coreference system. The covered phenomena are summarized in Table 1 and a detailed description is given in Appendix A.1.

Given a text t consisting of one or two consecutive sentences, each of our rules addresses a specific discourse phenomenon and has two parts: (a) conditions for matching the phenomenon in t, and (b) operations over a dependency tree annotated with coreference resolution. Applying the operations generates a fusion example ($x = (s_1, s_2), t$), in which (s_1, s_2) are two independent sentences originating from t, but stripped from the discourse phenomenon that tied them in t.

Figure 2 gives an example of a rule for the apposition structure. The rule is applied to the sentence *"The Jacksonville Jazz Piano Competition, a 30 year tradition, takes place at the Florida Theatre"*. First, the input is matched to the rule's condition. In this case, the condition is a single clause surrounded by two commas, which has a determiner as its first token and includes an apposition with incoming edge from a preceding token to the clause. Once matched, an example is generated. For this rule, the first sentence is created by removing the apposition clause, and the second sentence by removing the part after the clause and

inserting the appropriate *"be"* verb (*"is"*). Generation examples for all 9 rule types are provided in Table 1.

As explained in Section 1, solving sentence fusion involves more than just reverse-engineering the generation rules. The model needs to decide whether to insert a discourse connective with the right semantics, whether to merge the input sentences, and what syntactic construction (relative clause, coordination, apposition) is most appropriate in the given context.

Last, often several discourse phenomena occur in a single text t. Thus, we allow combining anaphora rules with one of the following rule types: discourse connective, inner connective and sentence coordination, which cover frequent combinations in our texts.

3.2 Building the DISCOFUSE Dataset

To create DISCOFUSE we retrieved the latest Wikipedia release and crawled the Web for several million sports articles. Documents were annotated with dependency trees and coreference resolution using Google Cloud Natural Language.[1]

We considered each sentence and pair of consecutive sentences in each document as candidates, applying the example generation process described in Section 3.1. Additionally, we added as examples sentence pairs from the original corpus that did not match any rule, that is (s_1, s_2) = t, so that a trained model would also learn when not to change the input. We filtered out examples with sentence length $\leq$ 6 tokens, and examples with non-ASCII characters. This process resulted in 44, 177, 443 sports examples and 16, 642, 323 Wikipedia examples. We randomly split these examples into 98% train, 1% dev, and 1% test sets,

---

[1] https://cloud.google.com/natural-language/

3445

Phenomenon	Example
Discourse connective	(A) Hebden Bridge is a popular place to live .
	(B) **However ,** space is limited due to the steep valleys and lack of flat land .
	(a) Hebden Bridge is a popular place to live .
	(b) Space is limited due to the steep valleys and lack of flat land .
Anaphora	(A) **Rider** entered the weekend averaging 23.0 points , good for 10th in the league .
	(B) **He** said those numbers mean little because of the Hawks ' 11 - 18 record.
	(a) **Rider** entered the weekend averaging 23.0 points , good for 10th in the league .
	(b) **Rider** said those numbers mean little because of the Hawks ' 11 - 18 record.
Forward connective	(A) **Although** the friendship somewhat healed years later , it was a devastating loss to Croly .
	(a) The friendship somewhat healed years later .
	(b) It was a devastating loss to Croly .
Inner connective	(A) Open workouts are held every Sunday **unless** the gym is closed for a holiday or other special events .
	(a) Open workouts are held every Sunday .
	(b) The gym is closed for a holiday or other special events .
Cataphora	(A) **Stating** that the proponents were unlikely to succeed in this appeal , **Walker** rejected the stay request on October 23 .
	(a) **Walker stated** that the proponents were unlikely to succeed in this appeal .
	(b) **Walker** rejected the stay request on October 23 .
Sentence coordination	(A) The time of the autumn floods came , **and** the hundred streams poured into the Yellow River .
	(a) The time of the autumn floods came .
	(b) The hundred streams poured into the Yellow River .
Verb phrase coordination	(A) The Sharks **started** the year 0 - 4 , **yet recovered** to claim sixth spot .
	(a) The Sharks **started** the year 0 - 4 .
	(b) The Sharks **recovered** to claim sixth spot .
Relative clause	(A) Kubler **, who** retired from cycling in 1957 **,** remained a revered figure in the wealthy alpine nation .
	(a) Kubler remained a revered figure in the wealthy alpine nation .
	(b) Kubler retired from cycling in 1957 .
Apposition	(A) The frigidarium **,** the last stop in the bathhouse **,** was where guests would cool off in a large pool .
	(a) The frigidarium was where guests would cool off in a large pool .
	(b) The frigidarium is the last stop in the bathhouse .

Table 1: Generated fusion examples for different phenomena. The input text is marked in uppercase blue, and the generated sentence pair is marked in lowercase red. We show in **boldface** parts that allow us to detect the target phenomenon.

making sure that each document contributes examples to only one of the split sets.

Like prior work (Malmi et al., 2018), we observed a skewed distribution of discourse phenomena in the data. Specifically, examples with anaphora or the connectives *"and"* and *"but"* constitute 99.7% of Sports and 59% of Wikipedia examples. Such a skewed distribution is likely to bias models and will fail to elucidate the ability of models to capture a wide range of linguistic phenomena. Therefore, we constructed a version of DISCOFUSE by down-sampling examples containing *"and"* and *"but"* or anaphora. The down-sampled dataset contains 12,080,513 Sports examples and 4,581,352 Wikipedia examples.

The resulting distributions of discourse types and most common connectives in the two parts of DISCOFUSE are provided in Appendix A.2. We will release both the original and the down-sampled versions of DISCOFUSE.

Rater Selection	SPORTS (%)	WIKIPEDIA (%)
Yes	83.4	86.0
No majority	10.9	8.9
No	5.7	5.1

Table 2: Rater evaluation understandability of the text after splitting. For each example, the majority of 5 raters was taken as the final rater selection.

4 DISCOFUSE Quality Evaluation

To assess the quality of the generated fusion examples in DISCOFUSE, we randomly selected 500 examples from each of the development sets of the Wikipedia and the Sports parts. We then conducted a crowdsourcing experiment in which each example was rated by 5 proficient English speakers, limiting each rater to at most 6 items. Each rater was presented with the two independent sentences in the example and was asked to indicate whether the text is understandable. If the rater answered "yes", she was then asked to characterize the relation between the sentences and how she would fuse them. We next detail the results.

Reason	Example
Original text unclear	(A) UPDATE: Peat falls because footwork and quickness. (a) UPDATE: Peat falls. (b) Footwork and quickness.
Missing context	(A) We were right on the heels of Spurs, although Everton were closing in. (a) We were right on the heels of Spurs. (b) Everton were closing in.
Bad rule generation	(A) He told reporters after the game his reaction was because he missed a wide-open Randall Cobb in the end zone. (a) He told reporters after the game his reaction was. (b) He missed a wide-open Randall Cobb in the end zone.

Table 3: Examples for three possible reasons for not understanding the text. In each example, (A) is the original text and (a) and (b) are the two sentences generated by our rules.

Error	Generated sentence
Extra comma	The space behind the fence in right field is blocked off **,** .
Missing determiner	*My* internist sent me for a mammogram and sonogram .
Bad pronoun replacement	**Lions** have a 3 - 1 record overall against Lions .

Table 4: Examples of grammatical errors introduced by our rules. The **red boldface** text was incorrectly inserted and the *blue italic* text was incorrectly removed.

4.1 Example Text Clarity

Raters were asked whether they can understand the text after the example is split. Table 2 summarizes this evaluation. Most examples were marked as understandable by the raters ("yes") – 86% of Wikipedia examples and 83.4% of Sports examples. The rest either had no majority of rater votes or were marked as not understandable.

To shed light on the possible reasons for obscurity, we analyzed 70 random examples that were not marked as understandable by the majority of raters. In 29 examples (41%) the original text was unclear and for 17 examples a broader context was needed (24%). In the remaining 24 examples (34%), our rules generated sentences with grammatical or semantic errors. Examples for these cases are in Table 3.

Additionally, we analyzed 100 random examples for grammatical errors, and found that our rules did not introduce any errors in 79 examples. For 15 examples, the errors did not modify the meaning of the text nor caused semantic errors. The detected grammatical errors include ex-

tra commas, missing determiners and bad pronoun replacements, and are demonstrated in Table 4.

4.2 Fusion Evaluation

Next, we evaluated agreement on the fusion task for the 847 examples marked as understandable in Section 4.1. Because there are many ways in which sentences can be fused, one cannot expect raters to produce the original text t verbatim. Instead, we analyzed three central decisions and estimated whether people agree on those: (a) whether to merge the two sentences into a single one or keep them separate; (b) whether there are entities in the text that should be replaced with nominal or pronominal anaphors or cataphors; and (c) which discourse connective to add (if any).

For the last question, we presented raters with one connective from each of the four coarse-grained senses for discourse connectives defined by the PDTB (Prasad et al., 2008): *comparison*, *expansion*, *contingency* and *temporal*, as well as a *no-connective* option. If the original text in the example includes a connective, we provided it as one of the options.

We observed a strong bias among raters towards refraining from performing any changes. E.g., while only 38% of the examples did not contain a connective in t, the raters chose not to add a connective in 69.2% of the cases. Similarly, only in 29.1% of the examples the two sentences were not merged into a single one, while the raters chose not to merge in 53.1% of the examples. Similar behavior was also observed by Malmi et al. (2018) and Rohde et al. (2016).

We further looked at the agreement between the rater majority and the 'gold' fusion decision. This analysis is shown in Table 5. Agreement on merging the input sentences into one is almost random (52%), since usually both options are valid. Consensus on whether to add an anaphor is higher, but not very high (63%), especially in sentences when the anaphor in t is a nominal rather than a pronoun. Finally, there is higher agreement on selecting the connective category (57%), for which the random baseline is 20%.

As mentioned, raters tend to keep the sentences unchanged. But in cases where raters agree to add a connective, agreement figures increase substantially. Specifically, when it is clear that a connective is needed, there is also high agreement for picking the right one (76%), for deciding whether

Decision Type	All Examples (%)	Examples with connectives (%)
Single / pair	52.0	70.1
Anaphor	63.4	70.1
Connective category	57.0	76.4
# Examples	847	271

Table 5: Average agreement for each fusion decision between the gold annotation and rater majority on examples marked as understandable by the raters. The right column considers only examples in which both the 'gold' and rater majority agreed that a connective should be added.

to add an anaphor (70%), and for deciding whether to merge the sentences or not (70%).

5 Supervised Neural Fusion Models

Using DISCOFUSE, we trained a Transformer seq2seq model (Vaswani et al., 2017) that reads the input sentence pair and generates a fused text. We report model performance on the test-set using automatic metrics as well as human evaluation. We also provide detailed analysis of the different phenomena captured by this model.

5.1 Experimental Settings

We tokenized all texts using byte-pair-encoding (Sennrich et al., 2015) and compared the following three Transformer models :

- **DFSPORT** - trained on the sports portion of DISCOFUSE after down-sampling.

- **DFWIKI** - trained on the Wikipedia portion of DISCOFUSE after down-sampling.

- **DFS+W** - trained on a 50%-50% mixture of the sports and Wikipedia portions of DISCO-FUSE after down-sampling.

All models share the same network architecture, based on the best practices discussed by Popel and Bojar (2018). We tuned parameters to select the best learning and dropout rates for each model with respect to the Exact Match objective (described in Section 5.2). Network architecture and hyper-parameters are in Appendix A.3. As a baseline, we also tested a model called COPY, which simply concatenates the two input sentences.

5.2 Automatic Evaluation Results

We evaluated model performance using two automatic metrics. The first is Exact Match (*Exact*) to see how often the model generates the exact same

SPORTS	Full		Sampled	
	SARI	*Exact*	*SARI*	*Exact*
DFSPORT	**81.9**	**42.3%**	**83.9**	**50.6%**
DFWIKI	77.8	31.7%	80.1	40.1%
DFS+W	80.7	38.3%	82.9	47.0%
COPY	40.0	1.1%	40.4	3.8%

WIKIPEDIA	Full		Sampled	
	SARI	*Exact*	*SARI*	*Exact*
DFSPORT	80.0	41.5%	80.0	41.9%
DFWIKI	**83.1**	**47.6%**	**84.5**	**51.1%**
DFS+W	82.8	46.7%	83.7	49.2%
COPY	40.3	1.0%	39.6	2.1%

Table 6: *Exact* and *SARI* scores of DFSPORT, DFWIKI, DFS+W and COPY, on the test sets of DISCOFUSE before (*Full*) and after down-sampling (*Sampled*).

Discourse phenomena	DFSPORT		DFWIKI	
	Exact	*SARI*	*Exact*	*SARI*
Apposition	94.8%	99.3	94.7%	99.6
Relative clause	84.3%	95.3	76.9%	92.6
Cataphora	79.9%	92.8	84.2%	95.8
Verb phrase coordination	58.4%	88.0	58.7%	88.8
None (control)	55.3%	73.2	54.2%	72.7
Anaphora	52.1%	83.7	47.7%	81.5
Inner connective	49.9%	83.9	51.6%	85.5
Sentence coordination	35.6%	80.9	31.7%	79.4
Inner connective + anaphora	32.6%	82.5	37.2%	83.7
Forward connective	27.5%	80.2	34.6%	82.8
Sentence coordination + anaphora	20.5%	80.9	16.3%	78.3
Discourse connective	14.2%	65.6	29.1%	73.4
Discourse connective + anaphora	2.5%	73.0	8.0%	72.1

Table 7: In-domain evaluation with breakdown by discourse phenomena. Performance of DFSPORT and DFWIKI on the sports and Wikipedia development sets.

text as the gold fusion. The second is *SARI* (Xu et al., 2016), which computes the set of added, removed, and kept n-grams in the model output, comparing the output both with the gold text and the input text. Then it computes the F_1 scores for these three sets and averages the scores. We compute *SARI* on up to 4-grams, as in Xu et al. (2016). We refrained from using metrics like BLEU because in fusion there is large overlap between the input sentences and their fused version, and such metrics do not capture well fine-grained differences of only a single word.

We note that our definition of *SARI*[2] slightly differs from the one given by Xu et al. (2016) in

---

[2]Our *SARI* implementation is available at: `https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/utils/sari_hook.py`

two aspects: (i) We define $\frac{0}{0} = 1$ when computing precision and recall, otherwise *SARI* could be less than 1 even if the output matches the gold text exactly. (ii) Instead of considering only the precision of deleted n-grams, we use F_1 for all three sets. Otherwise, *SARI* will give high scores to models that merely copy everything in the input, without even trying to infer what to delete.

Table 6 summarizes the results. When training and testing on the same domain, either Sports or Wikipedia, *SARI* score is a little above 80 points for the full dataset. Yet *Exact* is not high, around 42% for Sports and 47% for Wikipedia, showing that in the majority of the examples the model's fusion differs from the gold. Tested on the down-sampled test-set, performance increases significantly for *Exact*, especially on Sports, where discourse phenomena is more skewed.

We next turn to cross-domain evaluation. When applying a model trained on one domain to the other domain performance drops. This shows that the discourse phenomena distribution differs between the domains, indicating that transfer learning is not trivial even with these large datasets. This is especially evident when applying DFWIKI to Sports, where *Exact* falls from 42% to 32% on the full dataset and from 50% to 40% on the down-sampled one. Interestingly, when learning on the mixed training set, performance on both domains is close to in-domain performance, showing that the model has the capacity to handle both domains.

Finally, we take advantage of the provided annotation of the different discourse phenomena within each example in DISCOFUSE. We conducted a detailed analysis of in-domain model performance by discourse type, presented in Table 7. Results show that structural discourse types, such as apposition and relative clause, are easier to learn with both high exact match and *SARI* scores. While differences with respect to *SARI* scores are not large between phenomena, exact match varies more. Anaphora and verb phrase coordination are more challenging, but still require matching of the same noun (the named entity or the subject). On the other hand, discourse types that involve connective prediction, such as sentence coordination and discourse connective, require semantic understanding, and performance is significantly lower. In addition, when two discourse types are required for fusion, performance drops dramatically.

		# Examples	Detection
SPORTS	output = gold	525	50%
	output != gold	475	65%
	total	1000	**57%**
WIKIPEDIA	output = gold	528	50%
	output != gold	472	61%
	total	1000	**55%**

Table 8: Human detection (*Detection*) percentage for DFSPORT and DFWIKI on 1000 samples from each of the Sports and Wikipedia development sets. We report *Detection* for cases when model output differed from the gold, and cases when they were identical.

5.3 Human Evaluation Results

As our second experiment we employed crowd-sourcing to test how distinguishable the fusion model outputs are from the gold fused texts. Concretely, we present raters an independent sentence pair from DISCOFUSE and two fused versions - the gold version and one generated by a model. Raters were asked to detect the gold version. For each example, we took the majority of 5 raters as the final choice. This experiment mitigates the difficulties of automatic text generation evaluation, where many outputs are valid for a single input.

We sampled 1000 random examples from each development set of the two domains and applied the in-domain model to both. The raters were presented only with examples where the model output was different from the gold fusion, and assumed 50% detection accuracy otherwise.

Table 8 depicts the results. Out of cases when model output differed from the gold, raters were able to identify the human version in 65% of Sports examples and 61% of Wikipedia examples. Looking at the entire set, humans were able to identify the human version in 57% (Sports) and 55% (Wikipedia) of the cases. This shows that our Transformer model, applied over a dataset of millions of examples, is able to learn good fusions in general. Nevertheless, models are still far from perfect – human accuracy is clearly better than random and this improvement is statistically significant at a level of $p < 10^{-5}$ for Sports and $p < 10^{-3}$ for Wikipedia.

5.4 Alignment-based Analysis

We next present an analysis of the types of errors our models produce. To this end, we sampled 40K examples of DFSPORT and DFWIKI outputs on Sports and Wikipedia development sets. We then automatically aligned predicted sequences to gold

Connective	DFSPORT accuracy	DFWIKI accuracy	Top 3 connectives
and	50.9	53.7	and, but, ⟨other⟩
but	42.8	43.7	but, ⟨other⟩, and
because	61.5	60.7	because, ⟨other⟩, but
although	35.1	33.2	although, ⟨other⟩, but
so	50.6	50.2	so, but, and
or	70.5	72.1	or, and, ⟨other⟩
however	28.3	26.7	⟨other⟩, however, but
while	70.1	70.6	while, ⟨other⟩, but
so that	64.3	63.0	so that, ⟨other⟩, because
unless	68.9	67.0	unless, because, ⟨other⟩
for example	26.9	28.1	⟨other⟩, for example, however

Table 9: Alignment-based connective prediction accuracy for the most common connectives. When a model did not add a connective, the token ⟨other⟩ is used.
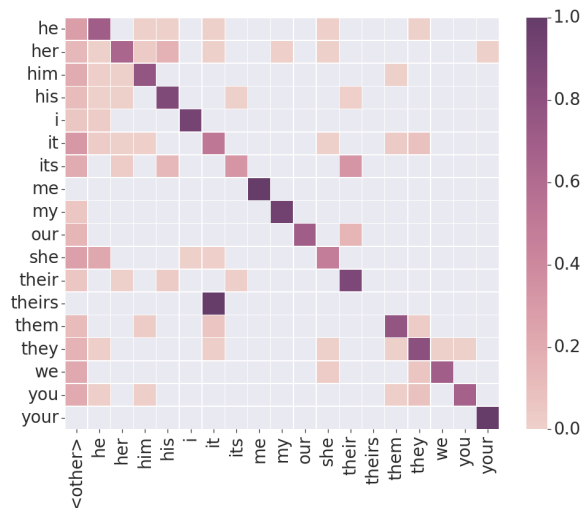

Figure 3: DFWIKI outputs versus the gold pronouns. Rows refer to gold pronouns and columns refer to aligned model outputs at the gold pronoun position. Values in each row are normalized to 1. Column ⟨other⟩ refers to model outputs that are not pronouns.

sequences and looked at the differences between aligned words. The trained models successfully learned to copy most of the input text, and thus errors due to alignment problems are rare.

We start by considering the semantic relation between the input sentences. Table 9 displays model accuracy in predicting the most common connectives in DISCOFUSE, as well as the top connectives predicted in this slot. We observe that when the model predicts a wrong connective, that connective is often reasonable, e.g., predicting *"but"* instead of *"and"* or *"however"*. Moreover, a second source of error is not adding a connective at all. It is also clear that some connectives, like *"however"*, *"although"* and *"for example"*, are harder to learn.

We also analyzed the models' ability to correctly infer pronoun anaphors including gender, possessive and plurality. Figure 3 shows the pronoun confusion matrix for DFWIKI,[3] where lines refer to gold pronouns and columns to the generated pronoun in the same position. The clear diagonal shows that in most cases, the model successfully outputs the correct pronoun. However, the ⟨other⟩ column indicates that occasionally the model tends not to replace the entity in the input with a pronoun anaphor. In addition, the model seems to struggle with possession and plural 3rd person (*"it"*, *"its"*, *"they"*, *"their"*, *"theirs"*).

---

[3] Results for DFSPORT are very similar.

6 Transfer Learning Experiment

With the DISCOFUSE approach we can collect a large amount of examples automatically. Still, these examples only reflect the manual rules that identify discourse phenomena. We wanted to see if DISCOFUSE covers enough cases such that a trained model would be helpful for testing on fusion datasets generated by different approaches.

6.1 Experimental Settings

In this experiment, we looked at the recently released WEBSPLIT dataset 1.0 (Narayan et al., 2017). It consists of examples $(t, \{s_i\}_{i=1}^n)$, where t is a sentence that verbalizes the same set of RDF triples as $\{s_i\}_{i=1}^n$. We note that WEBSPLIT was originally developed for sentence splitting, from t to $\{s_i\}_{i=1}^n$, but here we view its examples for the reverse fusion task: from $\{s_i\}_{i=1}^n$ to t. We only considered examples where $\{s_i\}_{i=1}^n$ corresponds to exactly two simpler sentences ($n = 2$). This leaves us with 135K training, 8K validation, and 8K test samples.

We tokenized the data using byte-pair-encoding (Sennrich et al., 2015) and compared three models: (i) The COPY baseline that concatenates the two input sentences, (ii) a model trained on WEBSPLIT alone, and (iii) a model pre-trained on DFWIKI and fine-tuned on WEBSPLIT.

For the last two models, we use the CopyNet architecture (Gu et al., 2016), which is similar

3450

Training data	SARI	Keep	Add	Delete
COPY	18.1	52.9	0.5	0.9
WEBSPLIT	40.5	44.6	7.8	**69.3**
DFWIKI + WEBSPLIT	**44.2**	**54.8**	**10.4**	67.5

Table 10: Fusion results on WEBSPLIT, measured by *SARI* and the F1 scores that compose it.

to state-of-the-art models for the splitting task on WEBSPLIT (Narayan et al., 2017; Botha et al., 2018). While the Transformer outperformed this model on our main experiments, here it overfit on the small training set of WEBSPLIT. The training details are provided in Appendix A.3.

6.2 Results

Table 10 shows the results of the experiment. Similarly to Section 5, we measured the model performance using *SARI*. Pre-training with DFWIKI improves *SARI* score by 9% compared to using WEBSPLIT alone. In particular, the F1 of the 'kept' and 'added' n-grams is significantly higher, by 23% and 33% respectively. Specifically, 'added' tokens refer also to correctly choosing discourse connectives, to which the large-scale examples in DISCOFUSE were likely helpful.

We note that even with pre-training, the SARI 'add' score is only 10.4. This is probably due to the large amount of paraphrasing done in WEBSPLIT, which makes it problematic for fusion evaluation (see also Section 2). For example:

> **Sentence 1:** *Bolt , a comic character AKA Larry Bolatinsky , was created by Paris Cullins and Ernie Colon .*
>
> **Sentence 2:** *Paris Cullins is a United States national .*
>
> **Gold:** *Larry Bolatinsky is the alternative name for the comic book character Bolt , which was created by Ernie Colon and the American Paris Cullins .*

Correctly inferring the added terms (shown in red) requires paraphrasing knowledge that is outside the scope of DISCOFUSE.

7 Conclusions

We presented DISCOFUSE, a large-scale dataset for sentence fusion that was generated by applying a rule-based method. It contains millions of examples from two domains, annotated with multiple discourse phenomena.

We used DISCOFUSE to build supervised neural models for sentence fusion and conducted fine-grained analyses of the results. Currently, our models fuse only two sentences together. We

would like to expand them to more input sentences in future work.

We also demonstrated DISCOFUSE's usefulness in a transfer learning setup on a different fusion test-set, hoping it would facilitate research on text fusion in data-scarce domains.

Acknowledgments

References

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724. Association for Computational Linguistics.

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. 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 1587–1597. Association for Computational Linguistics.

Jan A Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from wikipedia edit history. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yllias Chali, Moin Tanvee, and Mir Tafseer Nayeem. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 418–424.

José Camacho Collados. 2013. Splitting complex sentences for natural language processing applications: Building a simplified spanish corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472.

Micha Elsner and Deepak Santhanam. 2011. Learning to fuse disparate sentences. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 54–63. Association for Computational Linguistics.

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 322–330. Association for Computational Linguistics.

J. Gu, Z. Lu, H. Li, and V. O. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Association for Computational Linguistics (ACL)*.

Yutong Li, Nicholas Gekakis, Qiuze Wu, Boyue Li, Khyathi Chandu, and Eric Nyberg. 2018. Extraction meets abstraction: Ideal answer generation for biomedical questions. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 57–65.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *AAAI*, pages 2750–2756.

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic Prediction of Discourse Connectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient creation of an accurate sentence fusion corpus. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–320. Association for Computational Linguistics.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. 2008. The Penn discourse treebank 2.0. In *LREC*.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher NL Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Yiping Song, Rui Yan, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, and Dongyan Zhao. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *Proceedings of IJCAI-ECAI, 2018*.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1410–1418.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 290–297. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Rui Yan and Dongyan Zhao. 2018. Coupled context modeling for deep chit-chat: towards conversations between human and computer. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2574–2583. ACM.

A   Supplemental Material

A.1   Generation Rules

In this section we provide technical details of the generation rules used to create DISCOFUSE. For the sake of clarity, we provide a simplified version of the rules, that does not include edge cases and minor implementation details. The discourse connectives we considered in the rules were selected from the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and are listed in Table 12.

Given an input text, it is encoded with 3 lists: Z is the token list, Z_t is a list of POS tags, Z_l is a list of dependency labels (see Table 11). In addition, all entities mentioned in the text are extracted and stored such that for two token lists Z, R, the set $m(Z, R)$ holds all the mention pairs of the same entity in the two lists. Each rule is designed for a specific discourse phenomenon and contains two parts. First, a set of conditions is applied to the input lists to detect whether the phenomenon occurs in the text. If a discourse pattern has been identified, a short sequence of simple operations is applied to the input, yielding a new sentence pair. Table 13 summarizes the operations in use, which allow insertion and deletion of tokens and splitting of the input text.

Table 14 provides the technicalities of each rule, i.e. the detection conditions of the discourse structure, and the sequence of operations for generating a new sentence pair from it. A detailed example for two-rule execution process is given in Table 15.

As mentioned, the rules are simplified for clarity. However, we note two special cases where morphological modifications are required to produce text without grammatical errors. First, in some cases of forward connective and cataphora, the tense change of a verb is required when splitting the input sentence. For instance, in the cataphora example in Table 1, we change the verb *"stating"* to have a past tense – *"stated"*. Likewise, occasionally a *"be"* verb needs to be inserted when splitting a single sentence, as demonstrated in Figure 2. In our rules, we choose which *"be"* verb to insert based on the tense and perspective of the rest of the sentence.

A.2 DISCOFUSE Data Distribution

Figure 4 and Table 16 show the distributions of discourse types and most common connectives in the two parts of DISCOFUSE.

Analyzing the dataset reveals significant differences in discourse phenomena between the two types of documents (Figure 4). E.g., coordination is very common in Wikipedia while anaphora is dominant in Sports. Likewise, the distribution of discourse connectives is quite different (Table 16).

A.3 Neural Models Parameters

The models DFSPORT, DFWIKI, DFS+W share the same Transformer network architecture, that was originally proposed by Vaswani et al. (2017). During training, we split the samples to buckets

Notation	Definition
Token list Z	A list of tokens $\{z^{(1)}, ..., z^{(|Z|)}\}$
Z_t	The list of POS tags of Z, where $Z_t^{(i)}$ is the tag of $z^{(i)}$ for every $i = 1, ..., |Z|$.
Z_l	The list of dependency labels of Z, where $Z_l^{(i)}$ is the label of incoming edge of $z^{(i)}$ for every $i = 1, ..., |Z|$.
$m(Z, R)$	A set of mention pairs in Z, R: $\{\langle S_Z, S_R \rangle \mid S_R \prec R$ and $S_Z \prec Z$ are mentions of the same entity$\}$
$S \prec Z$	S is a span in Z, such that $\exists i \in 1, ..., |Z| - |S| + 1 : \forall j = 0, ..., |S| - 1 : s_{1+j} = z_{i+j}$
$S \sqsubseteq Z$	S is a prefix of Z, such that $\forall i = 1, ..., |S| : s_i = z_i$
$i \frown_U j$	There is an edge from the ith token to the jth token in the dependency tree of Z.
$\mathcal{C}_b$	A set of backward connectives.
$\mathcal{C}_s$	A set of intra-sentence connectives, which are either forward connectives or conjunctions.
$\mathcal{C}_f$	A set of forward connectives.
$\mathcal{C}_c$	A set of coordinating conjunctions.
$\mathcal{P}_r$	A set of relative pronouns.
$\mathcal{V}$	A set of POS tags for verbal phrases.

Table 11: Notation and definitions for Table 14 of generation rules. Z, R, S_Z, S_R are token lists and $1 \leq i, j \leq |Z|$ are indices. The full lists of connectives and POS tags are provided in Table 16.

Set	Values
$\mathcal{C}_b$	"accordingly", "additionally", "afterward", "alternatively", "although ,", "and", "as a result ,", "because of that", "because of this", "besides ,", "but", "by comparison ,", "by contrast ,", "by doing this ,", "by then ,", "consequently", "conversely", "else,", "finally ,", "for example", "for instance", "further ,", "furthermore", "hence ,", "however", "in contrast ,", "in fact ,", "in other words", "in particular ,", "in short ,", "in sum ,", "in the end ,", "in turn ,", "indeed ,", "instead ,", "lest", "likewise ,", "meantime ,", "in the meantime ,", "meanwhile ,", "moreover", "nevertheless", "next ,", "nonetheless", "on the contrary ,", "on the other hand ,", "or ,", "otherwise ,", "overall ,", "plus ,", "rather ,", "regardless ,", "similarly ,", "simultaneously ,", "specifically ,", "still ,", "then ,", "thereafter ,", "thereby ,", "therefore", "though ,", "thus ,", "ultimately ,", "whereas", "yet ,", "now ,", "second ,", "third ,", "basically ,", "this ,", "eventually ,", "obviously ,", "again ,", "fortunately ,", "luckily ,", "meaning ,", "interestingly ,", "anyway ,", "clearly ,"
$\mathcal{C}_s$	"because", ", because", "hence", ", while", "whereas", ", although", "although", "and although", "unless", "now that", ", now that", "so that", ", so that", "meaning", ", meaning"
$\mathcal{C}_f$	although, since, in addition to, aside from
$\mathcal{C}_c$	and, but, or, nor, yet, so, for
$\mathcal{P}_r$	who, which, whose, whom
$\mathcal{V}$	VB, VBD, VBG, VBN, VBP, VBZ

Table 12: Connectives and POS tags used in our detection rules. A preceding comma is allowed for conjunctions in $\mathcal{C}_c$. For the connectives *"although"* and *"since"* in $\mathcal{C}_f$, we do not allow a following comma.

Operation	Description
DELETE(X, i, n)	Delete a sequence of n tokens from X, starting from index i.
PREPEND(X, Y)	Attach the list Y at the beginning of X.
REPLACE(X, Y, Z)	Replace every occurrence of Y in X with Z, in a non-overlapping manner.
SPLIT(X, i)	Split X into two token lists $V = \{x_1, ..., x_{i-1}\}, W = \{x_i, ..., x_{|X|}\}$.
TRIM(X)	Delete all tokens in X after the first punctuation token, e.g. period, comma, etc.

Table 13: Operations upon token lists, which are used for generation of sentence pairs (Table 14). The arguments X, Y, Z are token lists and the arguments i, n are integers.

Phenomenon	Input	Detection	Generation
Discourse connective	(A,B)	$\exists S \in \mathcal{C}_b , i \in [1,5] : S \sqsubseteq \{b^{(i)}, ..., b^{(|B|)}\}$	$\textsc{Delete}(B, i, |S|)$
Anaphora	(A,B)	$m(B, A) \neq \emptyset$	$\textsc{Replace}(B, S_B, S_A)$ $\forall \langle S_B, S_A \rangle \in m(B, A)$
Forward connective	Z	$\exists S \in \mathcal{C}_f : S \sqsubseteq Z$ $\exists i : |S| + 1 < i < |Z| \wedge z^{(i)} = ","$	$A, B \leftarrow \textsc{Split}(Z, i)$ $\textsc{Delete}(A, 1, |S|)$ $\textsc{Delete}(B, 1, 1)$
Inner connective	Z	$\exists S \in \mathcal{C}_s : S \prec Z$	$A, B \leftarrow \textsc{Split}(Z, i)$ $\textsc{Delete}(B, 1, |S|)$ $\textsc{Trim}(B)$
Cataphora	Z	$\exists i : 1 < i < |Z| \wedge z^{(i)} = ","$ $Z_t^{(1)} = \text{VBG} \wedge Z_l^{(1)} = \text{vmod} \wedge Z_t^{(i+2)} \in \mathcal{V}$	$A, B \leftarrow \textsc{Split}(Z, i)$ $\textsc{Prepend}(A, \{b_1\})$ $\textsc{Delete}(B, 1, 1)$
Sentence coordination	Z	$\exists i < j \leq i + 5 : Z_l^{(i)} = \text{cc} \wedge Z_l^{(j)} = \text{conj}$ $\exists i < k < j : Z_l^{(k)} \in \{\text{nsubj, nsubjpass}\}$	$A, B \leftarrow \textsc{Split}(Z, i)$ $\textsc{Delete}(B, 1, 1)$
Verb phrase coordination	Z	$\exists i, j : i < j \leq i + 5 : Z_l^{(i)} = \text{cc} \wedge Z_l^{(j)} = \text{conj} \wedge Z_t^{(j)} \in \mathcal{V}$ $\exists k : k \curvearrowright_Z j \wedge Z_l^{(k)} = \text{root}$	$A, B \leftarrow \textsc{Split}(Z, i)$ $\textsc{Delete}(B, 1, 1)$ $\textsc{Prepend}(B, \{a_1, ..., a_{k-1}\})$
Relative clause	Z	$\exists i, j : 1 < i < j < |Z| \wedge z^{(i)} = z^{(j)} = ","$ $z^{(i+1)} \in \mathcal{P}_r$	$A, B \leftarrow \textsc{Split}(Z, j)$ $\textsc{Delete}(B, 1, 1)$ $\textsc{Delete}(A, i, 2)$ $\textsc{Prepend}(B, \{a_r, ..., a_{i-1}\})^*$
Apposition	Z	$\exists i, j : 1 < i < j < |Z| \wedge z^{(i)} = z^{(j)} = ","$ $Z_l^{(i+1)} \in \{\text{det, poss}\} \wedge \exists k : i < k < j, Z_l^{(k)} = \text{appos}$	$A, B \leftarrow \textsc{Split}(Z, j)$ $\textsc{Delete}(B, 1, 1)$ $\textsc{Delete}(A, i, 2)$ $\textsc{Prepend}(B, \{a_r, ..., a_{i-1}\})^*$

Table 14: Generation rules for sentence pairs. The rules apply for token lists Z, A, B, where Z represents a single sentence and A, B either represent two consecutive sentences or two consecutive sentence parts. *For the rules of relative clause and apposition, r is the index of the leftmost child in the dependency sub-tree of $a^{(i)}$.

---

1. Input
$Z = \{$Ruiz ordered his first shot to be retaken because Brazilian players entered the penalty area before his kick .$\}$

↓

2. Inner connective
Detection
For $S = $ "because" it holds that $S \in \mathcal{C}_s$ and $S \prec Z$.

Generation

$\textsc{Split}(Z, i)$	$A = \{$Ruiz ordered his first shot to be retaken .$\}$ $B = \{$Because Brazilian players entered the penalty area before his kick .$\}$
$\textsc{Delete}(B, 1, |S|)$	$B = \{$ Brazilian players entered the penalty area before his kick .$\}$
$\textsc{Trim}(B)$	*no effect at this case*

↓

3. Anaphora
Detection
$m(B, A) = \{(\text{his, Ruiz})\} \neq \emptyset$

Generation

$\textsc{Replace}$ $(B, \text{his, Ruiz})$	$B = \{$ Brazilian players entered the penalty area before Ruiz 's kick .$\}$

↓

4. Output sentence pair
$A = \{$Ruiz ordered his first shot to be retaken . $\}$
$B = \{$Brazilian players entered the penalty area before Ruiz 's kick . $\}$

Table 15: Detailed two-rule execution example. We show in red parts of the input that are used for detection or modified during execution. The input token list Z is of a single sentence. First, the rule for inner connective is applied, splitting Z into two sentences A, B, without the connective *"because"*. Then, applying the anaphora rule, the pronoun *"his"* in B is replaced with the entity it refers to in A, to obtain a new sentence pair.

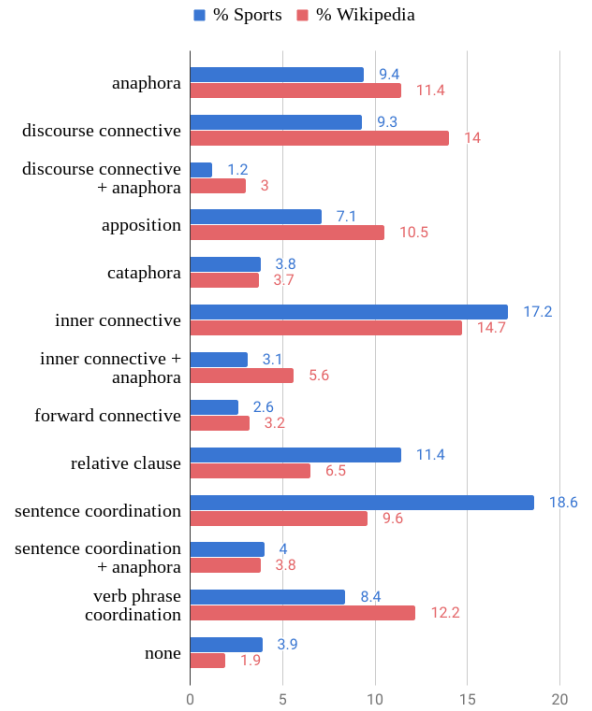by their text length, and use different batch size between 60-100 for each bucket. Further config-

Figure 4: Discourse type distribution of the sports and Wikipedia portions of DISCOFUSE after downsampling.

uration details and the hyperparameters used for training of each model are provided in Table 17.

In our transfer learning experiment, we trained two CopyNet models (Gu et al., 2016): a model

SPORTS		WIKIPEDIA	
	%		%
and	12.0	and	12.5
but	10.9	but	10.7
because	8.1	although	8.4
although	5.4	however	8.2
so	4.9	because	7.7
or	4.8	so that	2.1
however	3.5	while	2.0
while	2.4	or	1.8
so that	2.2	so	1.2
unless	2.2	for example	1.0

Table 16: Most common connectives in DISCOFUSE after down-sampling. Percentages are with respect to the entire dataset, including examples without a connective.

	DFSPORT	DFWIKI	DFS+W
number of hidden layers	7	7	7
hidden dimension	1024	1024	1024
filter size	2048	2048	2048
number of heads	16	16	16
beam width	4	4	4
attention dropout rate	0.1	0.2	0.2
ReLU dropout rate	0.4	0.3	0.2
learning rate	0.14	0.07	0.11

Table 17: Parameters and hyperparameters of the models DFSPORT, DFWIKI, DFS+W.

number of encoder layers	3
number of decoder layers	1
hidden dimension	128
beam width	20
scheduled sampling probability	0.2
dropout rate	0.2
learning rate	0.001
learning rate decay	0.98

Table 18: Parameters and hyperparameters of the CopyNet models used for transfer learning.

trained on WEBSPLIT alone, and a model pretrained on DFWIKI and finetuned on WEBSPLIT. The first model was trained for 200,000 steps on WEBSPLIT, whereas the second model was pretrained for 1 million steps on DFWIKI and then finetuned for 100,000 steps on WEBSPLIT. Again, the samples were split to buckets by their text length, with batch sizes between 25-125 for each bucket. The final test scores were computed with the parameters that maximize the validation *SARI* score during training. The network architecture and hyperparameters were shared between the models and not optimized during training. They are listed in Table 18.