# Semantically Driven Sentence Fusion: Modeling and Evaluation

**Eyal Ben-David[1], Orgad Keller[2], Eric Malmi[2], Idan Szpektor[2], Roi Reichart[1]**
[1]Faculty of Industrial Engineering and Management, Technion, IIT
[2]Google Research
eyalbd12@campus.technion.ac.il, roiri@technion.ac.il
{orgad,emalmi,szpektor}@google.com

## Abstract

Sentence fusion is the task of joining related sentences into coherent text. Current training and evaluation schemes for this task are based on single reference ground-truths and do not account for valid fusion variants. We show that this hinders models from robustly capturing the semantic relationship between input sentences. To alleviate this, we present an approach in which ground-truth solutions are automatically expanded into multiple references via curated equivalence classes of connective phrases. We apply this method to a large-scale dataset and use the augmented dataset for both model training and evaluation. To improve the learning of semantic representation using multiple references, we enrich the model with auxiliary discourse classification tasks under a multi-tasking framework. Our experiments highlight the improvements of our approach over state-of-the-art models. [1]

## 1 Introduction

Generative NLP tasks, such as machine translation and summarization, often rely on human generated ground truth. Datasets for such tasks typically contain only a single reference per example. This may result from the costly effort of human annotations, or from collection methodologies that are restricted to single reference resources (e.g., utilizing existing corpora; Koehn, 2005; Nallapati et al., 2016). However, typically there are other possible generation results, such as ground-truth paraphrases, that are also valid. Failing to consider multiple references hurts the development of generative models, since such models are considered correct, at both training and evaluation, only if they follow one specific and often arbitrary generation path per example.

In this work we address *Sentence Fusion*, a challenging task where a model should combine related sentences, which may overlap in content, into a compact coherent text. The output should preserve the information in the input sentences as well as their semantic relationship. It is a crucial component in many NLP applications, including text summarization, question answering and retrieval-based dialogues (Jing and McKeown, 2000; Barzilay and McKeown, 2005; Marsi and Krahmer, 2005; Lebanoff et al., 2019; Szpektor et al., 2020).

Our analysis of state-of-the-art fusion models (Geva et al., 2019; Rothe et al., 2019) indicates that they still struggle to correctly detect the semantic relationship between the input sentences, which is reflected in inappropriate discourse marker selection in the generated fusions (§4). At the same time, DISCOFUSE (Geva et al., 2019), the dataset they use, is limited to a single reference per example, ignoring discourse marker synonyms such as 'but' and 'however'. Noticing that humans tend to judge these synonyms as equally suitable (§3), we hypothesize that relying on single references may limit the performance of those models.

To overcome this limitation, we explore an approach in which ground-truth solutions are automatically expanded into multiple references. Concretely, connective terms in gold fusions are replaced with equivalent terms (e.g., {'*however*', '*but*'} ), where the semantically equivalent sets are derived from the Penn Discourse TreeBank 2.0 (Prasad et al., 2008). Human evaluation of a sample of these generated references indicates the high quality of this process (§3). We apply our method to automatically augment the DISCOFUSE dataset with multiple references, using the new dataset both for evaluation and training. We will make this dataset publicly available.

We then adapt a seq2seq fusion model in two ways so that it can exploit the multiple references in the new dataset (§4). First, each training example with its multiple references is converted into

[1]Our code is at https://github.com/eyalbd2/Semantically-Driven-Sentence-Fusion.

multiple examples, each consisting of the input sentence pair with a different single reference fusion. Hence, the model is exposed to a more diverse and balanced set of fusion examples. Second, we direct the model to learn a common semantic representation for equivalent surface forms offered by the multiple references. To that end, we enhance the model with two auxiliary tasks: Predicting the type of the discourse relation and predicting the connective pertaining to the fused output, as derived from the reference augmentation process.

We evaluate our model against state-of-the-art models in two experimental settings (§5, 6): In-domain and cross-domain learning. The cross-domain setting is more challenging but may also be more realistic as labeled data is available only for the source domain but not for the target domain. To evaluate against multiple-reference examples, we measure the similarity of each generated fusion to each of the ground-truth fusions and report the highest score. This offers a more robust analysis, and reveals that the performance of fusion models is higher than previously estimated. In both settings, our model demonstrates substantial performance improvement over the baselines.

## 2 Related Work

### 2.1 Fusion Tasks

Traditionally, supervised sentence fusion models had access to only small labeled datasets. Therefore, they relied on hand-crafted features (Barzilay and McKeown, 2005; Filippova and Strube, 2008; Elsner and Santhanam, 2011; Filippova, 2010; Thadani and McKeown, 2013). Recently, DISCOFUSE, a large-scale dataset for sentence fusion, was introduced by Geva et al. (2019). This dataset was generated by automatically applying hand-crafted rules for 12 different discourse phenomena to break fused text examples from two domains, Wikipedia and Sports news, into two unfused sentences, while the content of the original text is preserved. We follow prior work (Malmi et al., 2019; Rothe et al., 2019) and use the balanced version of DISCOFUSE, containing ∼16.5 million examples, where the most frequent discourse phenomena were down-sampled.

With DISCOFUSE, it became possible to train data-hungry neural fusion models. Geva et al. (2019) showed that a Transformer model (Vaswani et al., 2017) outperforms an LSTM-based (Hochreiter and Schmidhuber, 1997) seq2seq model on this dataset. Malmi et al. (2019) further improved accuracy by introducing LaserTagger, modeling sentence fusion as a sequence tagging problem. Rothe et al. (2019) set the state-of-the-art with a BERT-based (Devlin et al., 2019) model.

Related to sentence fusion is the task of predicting the discourse marker that should connect two input sentences (Elhadad and McKeown, 1990; Grote and Stede, 1998; Malmi et al., 2018). It is typically utilized as an intermediate step to improve downstream tasks, mainly for discourse relation prediction (Pitler et al., 2008; Zhou et al., 2010; Braud and Denis, 2016; Qin et al., 2017). Connective prediction was included in multi-task frameworks for discourse relation prediction (Liu et al., 2016) and unsupervised sentence embedding (Jernite et al., 2017; Nie et al., 2019). We follow this approach of guiding a main task with the semantic information encompassed in discourse markers, studying it in the context of sentence fusion.

### 2.2 Generation Evaluation

Two main approaches are used to evaluate generation models against a single gold-truth reference. The first estimates the correctness of a generated text using a 'softer' similarity metric between the text and the reference instead of exact matching. Earlier metrics like BLEU and ROUGE (Papineni et al., 2002; Lin, 2004), considered n-gram agreement. Later metrics matched words in the two texts using their word embeddings (Lo, 2017; Clark et al., 2019). More recently, contextual similarity measures were devised for this purpose (Lo, 2019; Wieting et al., 2019; Zhao et al., 2019; Zhang et al., 2020; Sellam et al., 2020). In §7 we provide a qualitative analysis for the latter, presenting typical evaluation mistakes made by a recently-proposed contextual-similarity based metric (Zhang et al., 2020). This analysis reveals properties that characterize such methods, which make them less suitable for our task.

The second approach extends the (single) reference into multiple ones, by automatically generating paraphrases of the reference (a.k.a pseudo-references) (Albrecht and Hwa, 2008; Yoshimura et al., 2019; Kauchak and Barzilay, 2006; Edunov et al., 2018; Gao et al., 2020). Our method (§3.3) follows this paradigm. It applies curated paraphrase rules to generate highly accurate variations, putting an emphasis on precision. This is opposite to the recall-oriented similarity-based approaches,

which may detect correct fusions beyond paraphrasing approaches, but may also promote erroneous solutions due to their soft matching nature.

## 3 Multiple References in Sentence Fusion

In this section we discuss the limitations of using single references for evaluation and training in sentence fusion. We then propose an automatic, precision-oriented method to create valid fusion variants. Human-based evaluation confirms the reliability of our method, which generates pseudo-references that are considered as suitable as the original references. Finally, we demonstrate the effectiveness of the new references for fusion evaluation. Our observations, which are used here for reference generation and evaluation, will also guide our fusion model design and training (§4).

### 3.1 Single-reference Based Evaluation

Recent fusion works (Geva et al., 2019; Malmi et al., 2019; Rothe et al., 2019) rely on single references for training and evaluation. Two evaluation metrics are used: (1) EXACT, where the generated fusion should match the reference exactly, and (2) SARI (Xu et al., 2016), which measures the F1 score of kept and added n-grams, and the precision of deleted n-grams, compared to the gold fusion and the input sentences, weighting each equally.

A significant limitation of the above metrics, when measured using a single fusion reference, is that they do not properly handle semantically equivalent variants. For EXACT this is obvious, since even one word difference would account as an error. In SARI, the penalty for equivalent words, e.g., '*but*' and '*however*', and non-equivalent ones, e.g., '*but*' and '*moreover*', is identical.

To validate this, we conducted a qualitative single-reference evaluation of a fusion model (AuxBERT, §4.3) under the EXACT metric. We randomly selected 50 examples, assessed as mistakes, from the dev sets of both DISCOFUSE's domains. Analyzing these mistakes, we identified six types of errors (Table 1).

The distribution of these error types is depicted in Table 2. We note that the most frequent error type refers to *valid* fusion variants that differ from the gold fusion: As much as 40% and 44% of the examples in the Wikipedia and the Sports datasets, respectively. While the sample size is too small for establishing accurate statistics, the high-level trend is clear, indicating that a significant portion of the generated fusions classified as mistakes by the EXACT metric are in fact correct.

A possible solution would be to rely on single references, but use 'softer' evaluation metrics (see §2.2). We experimented with the state-of-the-art BERTScore metric (Zhang et al., 2020) and noticed that it often fails to correctly account for the semantics of discourse markers (see §7), which is particularly important for sentence fusion. Furthermore, we notice that recent soft metrics depend on trainable models, mainly BERT (Devlin et al., 2019), which is also used in state-of-the-art fusion models (Malmi et al., 2019; Rothe et al., 2019). Thus, we expect these metrics to struggle in evaluation when fusion models struggle in prediction.

### 3.2 Multi-Reference Generation

Generation of valid variants that differ from the ground-truth reference is a challenge for various generation tasks. For open-ended tasks like text summarization, researchers often resort to manually annotating multiple valid reference summaries for a small sample of examples. Sentence fusion, however, is a more restricted task, enabling high-quality automatic paraphrasing of gold fusions into multiple valid references. We introduce a precision-oriented approach for this aim.

Instead of generating arbitrary semantically equivalent paraphrases, we focus on generating variants that differ only by discourse markers, which are key phrases to be added when fusing sentences. The *Penn Discourse TreeBank 2.0* (PDTB; Prasad et al., 2008) contains examples of argument pairs with an explicit discourse marker and a human-annotated sense tag. The same marker may be associated with multiple sense tags (for instance, *since* may indicate both temporal and causal relations), and for our precision-oriented approach we only considered unambiguous markers.

Concretely, we identified three PDTB sense tags most relevant to the DiscoFuse dataset and chose the markers whose tag is one of those three in at least 90% of all PDTB argument pairs with an explicit marker. The resulting clusters are presented in Table 3.[2] Finally, we add a fourth cluster containing relative clause paraphrases (such as *who is*

[2]Some connective occurrences differ only in the addition or omission of a punctuation mark, e.g., '*but*' and '*but,*'. From a sample of examples, we did not find cases in which punctuation changes the semantics of the resulting connection. Therefore, for every connective phrase in Table 3, we automatically consider also its variants that differ only in punctuation.

| Mistake type | Examples |
|---|---|
| Correct fusion variant | (a+b) It is situated around Bad Segeberg but not part of it . Bad Segeberg is the seat of the AMT .<br>(I) It is situated around Bad Segeberg **,** the seat of the AMT **,** but not part of it .<br>(G) It is situated around Bad Segeberg **, which is** the seat of the AMT **,** but not part of it . |
| Missing/added anaphora | (a+b) Of the three , purple is preferred . Purple reinforces the red .<br>(I) Of the three , purple is preferred **because** purple reinforces the red .<br>(G) Of the three , purple is preferred **because it** reinforces the red . |
| Missing context/info | (a+b) Bolger quickly defeated Mclay . Gair himself took the position of deputy leader .<br>(I) Bolger quickly defeated Mclay **, while** Gair himself took the position of deputy leader .<br>(G) Bolger quickly defeated Mclay **, and** Gair himself took the position of deputy leader . |
| Missing/added punctuation | (a+b) Gair was born in Dunedin . Gair was moved to Wellington when young .<br>(I) Gair was born in Dunedin **but** moved to Wellington when young .<br>(G) Gair was born in Dunedin **, but** moved to Wellington when young . |
| Annotation error | (a+b) Krishnamurti 's notebook . By Krishnamurti , Krishnamurti ( hardcover ) .<br>(I) Krishnamurti ' s notebook . By krishnamurti ( hardcover ) .<br>(G) Krishnamurti ' s notebook . By Jiddu , krishnamurti ( hardcover ) . |
| Semantic Errors | (a+b) George Every never married . George Every never had children.<br>(I) George Every never married **or** had children .<br>(G) George Every never married **nor** had children . |

Table 1: Examples of various model error types. The input text is marked with a+b, the generated fusion is marked with I and the ground-truth fusion is marked with G. Errors are highlighted in bold font.

| Error Type | Wikipedia | Sports |
|---|---|---|
| Correct Fusion Variant | 40% | 44% |
| Miss/Add Anaphora | 2% | 2% |
| Missing Context | 18% | 18% |
| Miss/Add Punctuation | 22% | 18% |
| Annotation Error | 8% | 8% |
| Semantic Error | 10% | 10% |

Table 2: Error type distribution in 50 dev examples.

| **Cause** | **Conjunction** | **Comparison** |
|---|---|---|
| As a result | Furthermore | However |
| Hence | And | Yet |
| Consequently | Additionally | Still |
| Thus | Moreover | Nevertheless |
| Therefore | Plus | Although |
| | | But |

Table 3: Clusters of interchangeable connective markers constructed based on PDTB 2.0 sense tags.

*the* and *which is a*). Paraphrases from this cluster are not equivalent and cannot be replaced one with each other. Instead, they are replaceable with apposition paraphrases (as demonstrated in Table 4, under *Relative Clause*).

Given a DISCOFUSE target fusion $t_i$, if $t_i$ is annotated with a connective phrase $c_i$ that appears in one of our semantic clusters, we define the set $\mathcal{V}(t_i)$ that includes $t_i$ and its variants. These variants are automatically generated by replacing the occurrence of $c_i$ in $t_i$ by the other cluster members. Table 4 demonstrates this process. More details and examples are in the appendix (§A).

### 3.3 The Quality of Multiple References

To validate the reliability of our automatically generated variants as ground-truth fusions, we evaluate their quality with human annotators. To this end, we randomly sampled 350 examples from the DISCOFUSE dev sets (Wikipedia and Sports). Each example consists of two input sentences, and two fusion outcomes: the gold fusion and one automatically generated variant. We then conducted—using Amazon Mechanical Turk—a crowd-sourcing experiment where each example was rated by 5 native English speakers. Each rater indicated if one fusion outcome is better than the other, or if both outcomes are of similar quality (good or bad). We considered the majority of rater votes for each example. Table 5 summarizes this experiment. It shows that the raters did not favor a specific fusion outcome, which reinforces our confidence in our precision-based automatic generation method.

To demonstrate the benefit of our generated multiple references in fusion evaluation, we re-evaluated the mistakes marked by single-reference EXACT in §3.1. Concretely, each gold fusion $t_i$ was automatically expanded into the multiple reference set $\mathcal{V}(t_i)$. We define a multi-reference accuracy: MR-EXACT $= 1/N \sum_{i=1}^{N} \max_{t \in \mathcal{V}(t_i)} \mathbb{1}[f_i = t]$, where $f_i$ is the generated fusion for example $i$, $\mathbb{1}$ is the indicator function, and $N$ is the size of the test-set. MR-EXACT[3] considers a generated fusion for an example correct if it matches one of the

[3]We also define the MR-SARI measure. It follows MR-EXACT's formulation, taking the maximum over the SARI score between the generated fusion and the references.

| Phenomenon | | Examples |
|---|---|---|
| Conjunction | G | It'll work because god says so . **Plus ,** we are both willing to fight for it . |
| | V | It'll work because god says so . **Furthermore ,** we are both willing to fight for it . |
| | V | It'll work because god says so **, and** we are both willing to fight for it . |
| Cause | G | But the client is on a break. **Therefore** I'm on a break. |
| | V | But the client is on a break. **Hence** I'm on a break. |
| Comparison | G | It might sound like a nightmare **but** this news made this day one of the greatest of my life . |
| | V | It might sound like a nightmare **. Yet ,** this news made this day one of the greatest of my life . |
| Relative Clause | G | She is famed for her noble art Raikiri, **which is a** slash powered by lightning, that is believed to be inevitable. |
| | V | She is famed for her noble art Raikiri, **a** slash powered by lightning, that is believed to be inevitable. |

Table 4: Examples of automatic variant generation for fusion phenomena. The gold fusion is marked by G. The automatically generated variants are marked by V. Parts that were changed during variant generation are **boldfaced**.

| Rating Type | Preference (%) |
|---|---|
| Both equally good | 74.6 |
| Original fusion is better | 7.1 |
| Generated variant is better | 9.4 |
| Both equally bad | 2.3 |
| No majority | 6.6 |

Table 5: Raters' preferences when comparing original DISCOFUSE fusions to fusions generated by our automatic augmentation process.

variants in $\mathcal{V}(t_i)$. We measured an absolute error reduction of 15% in both domains, where all these cases come from the *correct fusion* class of Table 2.

## 4 A Semantically Directed Fusion Model

In the previous section we have established the importance of multiple references for sentence fusion. We next show (§4.1) that the state-of-the-art fusion model fails to detect the semantic relationship between the input sentences. We aim to solve this problem by expanding the training data to include multiple-references (MRs) per input example, where together these references provide a good coverage of the semantic relationship and are not limited to a single connective phrase. We then present our model (§4.3), which utilizes auxiliary tasks in order to facilitate the learning of the semantic relationships from the MR training data (§4.2).

### 4.1 The SotA Model: Error Analysis

Rothe et al. (2019) set the current state-of-the-art results on the DISCOFUSE dataset with a model that consists of a pre-trained BERT encoder paired with a randomly initialized Transfomer decoder, which are then fine-tuned for the fusion task. We re-implemented this model, denoted here by BERT, which serves as our baseline. We then evaluated BERT on DISCOFUSE using MR-EXACT (§3.3) and report its performance on each of the discourse phenomena manifested in the dataset (Table 11).

We found that this model excels in fusion cases that are *entity-centric* in nature. In these cases, the fused elements are different information pieces related to a specific entity, such as pronominalization and apposition (bottom part of Table 11). These fusion cases do not revolve around the semantic relationship between the two sentences. This is in line with recent work that has shown that the pre-trained BERT captures the syntactic structure of its input text (Tenney et al., 2019).

On the other hand, the performance of the BERT model degrades when fusion requires the detection of the *semantic relationship* between the input sentences, which is usually reflected via an insertion of a discourse marker. Indeed, this model often fails to identify the correct discourse marker (top part of Table 11). Table 6 demonstrates some of the semantic errors made by BERT.

### 4.2 Automatic Dataset Augmentation

We aim to expose a fusion model to various manifestations of the semantic relation between the input sentences in each training example, rather than to a single one, as well as to reduce the skewness in connective occurrences. We hypothesize that this should help the model better capture the semantic relationship between input sentences.

To this end, we utilize our implementation of the variant set $\mathcal{V}$ (§3.2). Specifically, for each training example $(s_i^1, s_i^2, t_i)$, we include the instances $\{(s_i^1, s_i^2, t') \mid t' \in \mathcal{V}(t_i)\}$ to the augmented training set. We then train a fusion model on this augmented dataset. The augmented dataset balances between variants of the same fusion phenomenon because if in the original dataset one variant was prominent, its occurrences are now augmented with occurrences of all other variants that can be offered by $\mathcal{V}$. We denote the baseline model trained on the augmented dataset by AugBERT.

| | Examples |
|---|---|
| (I) | Grace is told she can not get pregnant **and** IVF is unlikely to help. |
| (G) | Grace is told she can not get pregnant **because** IVF is unlikely to help. |
| (I) | The mounds now appear smaller than they did in the past **because** extensive flooding in the centuries since their construction has deposited 3 feet. |
| (G) | The mounds now appear smaller than they did in the past **, although** extensive flooding in the centuries since their construction has deposited 3 feet. |
| (I) | A Grand Compounder was a degree candidate at the University of Oxford who was required to pay extra for his degree **because** he had a certain high level of income. |
| (G) | A Grand Compounder was a degree candidate at the University of Oxford who was required to pay extra for his degree **unless** he had a certain high level of income. |
| (I) | The Battalion lost 41 men killed **or** died of wounds received on 1 July 1916. |
| (G) | The Battalion lost 41 men killed **and** died of wounds received on 1 July 1916. |

Table 6: Examples of semantic errors made by the BERT model. The generated fusion is marked with I and the ground-truth fusion is marked with G. These examples are handled correctly by our AuxBERT model.

### 4.3 Semantically Directed Modeling

Multiple references introduce diversity to the training set that could guide a model towards a more robust semantic representation. Yet, we expect that more semantic directives would be needed to utilize this data appropriately. Specifically, we hypothesize that the lower performance of the state-of-the-art BERT on semantic phenomena is partly due to its *mean cross-entropy* (MCE) loss function:

$$\ell_{\text{gen}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\log p(t_{i,j}|s_i^1, s_i^2, t_{i,1..j-1})$$

where $N$ is the size of the training-set, $T_i$ is the length of the target fusion $t_i$, $t_{i,j}$ is the $j$-th token of $t_i$, and $p(w|s_i^1, s_i^2, \text{pre})$ is the model's probability for the next token to be $w$, given the input sentences $s_i^1, s_i^2$ and the previously generated prefix pre.

As discussed earlier, the word-level overlap between the fusion and its input sentences is often high. Hence, many token-level predictions made by the model are mere copies of the input, and should be relatively easy to generate compared to new words that do not appear in the input. However, as the MCE loss does not distinguish copied words from newly generated ones, it would incur only a small penalty if only one or two words in a long fused sentence are incorrect, even if these words form an erroneous discourse marker. Moreover, the loss function does not directly account for the semantic (dis)similarities between connective terms. This may misguide the model to differentiate between similar connective terms, such as '*moreover*' and '*additionally*'.

To address these problems, we introduce a multi-task framework, where the main fusion task is jointly learned with two auxiliary classification tasks, whose goal is to make the model explicitly consider the semantic choices required for correct fusion. The first auxiliary task predicts the type of discourse phenomenon that constitutes the fusion act out of 12 possible tags (e.g. *apposition* or *discourse connective*; see Table 11). The second auxiliary task predicts the correct connective phrase (e.g. '*however*', '*plus*' or '*hence*') out of a list of 71 connective phrases. As gold labels for these tasks we utilize the structured information provided for each DISCOFUSE example, which includes the ground-truth discourse phenomenon and the connective phrase that was removed as part of the example construction. We denote this model AuxBERT and our full model with auxiliary tasks trained on the multiple-reference dataset AugAuxBERT.

The AuxBERT architecture is depicted in Figure 1. Both the fusion task and the two auxiliary classification tasks share the contextualized representation provided by the BERT encoder. Each classification task has its own output head, while the fusion task is performed via a Transformer decoder. The token-level outputs of the BERT encoder are processed by the attention mechanism of the Transformer decoder. BERT's CLS token, which provides a sentence-level representation, is post-processed by the *pooler* (following Devlin et al., 2019) and is fed to the two classification heads. The fusion component of the model is identical to Rothe et al. (2019) (BERT).

The three tasks we employ are combined in the following objective function:

$$\ell_{\text{total}} = \ell_{\text{gen}} + \alpha \cdot \ell_{\text{type}} + \beta \cdot \ell_{\text{conn}}$$

where $\ell_{\text{gen}}$ is the cross-entropy loss of the generation task, while $\ell_{\text{type}}$ and $\ell_{\text{conn}}$ are the cross-entropy losses of the discourse type and connective phrase predictions, respectively, with scalar

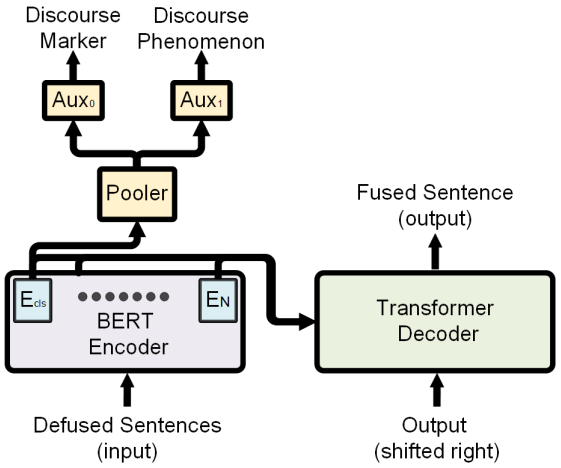


Figure 1: The AuxBERT architecture. $Aux_0$ and $Aux_1$ are classification layers for our auxiliary tasks.

weights $\alpha$ and $\beta$. We utilize a pre-trained BERT encoder and fine-tune only its top two layers.

## 5 Experimental Setup

We follow prior work and use the balanced version of DISCOFUSE (§2). The dataset consists of 4.5M examples for Wikipedia ('W') and 12M examples for Sports ('S'), split to 98% training, 1% test and 1% dev. We evaluate fusion models both in in-domain and cross-domain settings (training in one domain and testing on the other domain). We denote with W $\rightarrow$ S the setup where training is done on Wikipedia and testing on Sports, and similarly use S $\rightarrow$ W for the other way around.

We evaluate the following fusion models:[4]

**Transformer** - the Transformer-based model by Geva et al. (2019).

**LaserTagger** - the sequence tagging model by Malmi et al. (2019).

**BERT** - the BERT-based state-of-the-art model by Rothe et al. (2019).

**AugBERT** - BERT trained on our augmented MR training set (§4.2).

**AuxBERT** - Our multitask model (§4.3).

**AugAuxBERT** - Our multitask model trained on our augmented MR training set (§4.3).

All baselines used the same parameter settings described in the cited works, and our models follow the parameter settings in Rothe et al. (2019).

[4]Relevant code URLs are in the supplementary material.

| | W | W $\rightarrow$ S | S | S $\rightarrow$ W |
|---|---|---|---|---|
| LaserTagger | 56.1 | 51.2 | 59.7 | 51.2 |
| BERT | 67.9 | 57.4 | 63.2 | 59.7 |
| AugBERT | 69.9 | 59.7 | 65.7 | 62.7 |
| AuxBERT | 68.5 | 58.1 | 64.2 | 61.2 |
| AugAuxBERT | **71.0** | **60.9** | **67.1** | **63.9** |

Table 7: Multi reference EXACT (MR-EXACT) results.

| | W | W $\rightarrow$ S | S | S $\rightarrow$ W |
|---|---|---|---|---|
| LaserTagger | 79.8 | 79.5 | 82.7 | 77.6 |
| BERT | 90.3 | 86.4 | 88.0 | 86.6 |
| AugBERT | 90.7 | 86.5 | 88.7 | 86.8 |
| AuxBERT | 90.6 | **87.0** | 88.4 | 86.7 |
| AugAuxBERT | **91.1** | **87.0** | **89.2** | **87.0** |

Table 8: Multi reference SARI (MR-SARI) results.

| | W | W $\rightarrow$ S | S | S $\rightarrow$ W |
|---|---|---|---|---|
| Transformer | 51.1 | 40.1 | 50.6 | 41.9 |
| LaserTagger | 54.6 | 49.8 | 58.4 | 49.7 |
| BERT | 63.9 | 55.5 | 60.6 | 55.9 |
| AugBERT | 53.0 | 46.5 | 51.7 | 46.2 |
| AuxBERT | **65.0** | **56.5** | **61.8** | **57.0** |
| AugAuxBERT | 53.7 | 47.7 | 52.9 | 47.3 |

Table 9: Single reference EXACT results.

| | W | W $\rightarrow$ S | S | S $\rightarrow$ W |
|---|---|---|---|---|
| Transformer | 84.5 | 80.1 | 83.9 | 80.0 |
| LaserTagger | 79.1 | 78.6 | 81.9 | 76.4 |
| BERT | 89.2 | 85.8 | 87.2 | 85.3 |
| AugBERT | 85.3 | 81.9 | 83.9 | 82.5 |
| AuxBERT | **89.5** | **86.0** | **87.6** | **85.5** |
| AugAuxBERT | 85.7 | 82.5 | 84.4 | 82.9 |

Table 10: Single reference SARI results.

The same batch size and number of training steps were used in all models, thus training on the same number of examples when using either the original DISCOFUSE or our augmented version. The $\alpha$ and $\beta$ hyper-parameters of the multi-task objective are tuned on the dev set (see the supp. material).

## 6 Results

We report results with MR-EXACT (Table 7) and MR-SARI (Table 8). To maintain compatibility with prior work, we also report results with single reference (SR) EXACT (Table 9) and SARI (Table 10). Boldfaced figures in the tables are statistically significant with $p < 0.001$ compared to the second best model (using McNemar's paired test for EXACT and the Wilcoxon signed-rank test for SARI (Dror et al., 2018)).

For the SR evaluation, our AuxBERT is best performing, indicating the value of our multitask framework. On the other hand, training with the augmented dataset has a negative impact. This is

| Discourse phenomena | BERT S | BERT W | AugAuxBERT S | AugAuxBERT W |
|---|---|---|---|---|
| VP coordination | 68.5 | 67.1 | 67.9 | 67.8 |
| Inner connective | 66.0 | 71.3 | **69.3** | **74.4** |
| Inner connective+A | 46.0 | 58.0 | **52.6** | **61.0** |
| Sentence coordination | 52.1 | 56.4 | **59.4** | **63.5** |
| Sentence coordination+ A | 32.0 | 40.1 | **42.1** | **48.6** |
| Forward connective | 61.7 | 82.8 | **67.2** | 81.9 |
| Discourse connective | 29.6 | 49.0 | **48.0** | **61.6** |
| Discourse connective+A | 5.3 | 18.5 | **22.7** | **30.6** |
| Total Semantic | 52.7 | 59.5 | **59.7** | **64.9** |
| Apposition | 98.4 | 98.0 | 98.6 | 98.6 |
| Relative Clause | 90.9 | 91.1 | 91.9 | 89.4 |
| Cataphora | 91.5 | 94.0 | 90.6 | 94.0 |
| None | **66.9** | 68.1 | 57.6 | 68.5 |
| Anaphora | **62.0** | 62.5 | 58.4 | 61.9 |
| Total Entity-centric | **82.5** | 83.0 | 80.6 | 82.8 |

Table 11: In-domain evaluation with MR-EXACT, split by fusion phenomena. Boldfaced figures represent big gaps (more than 1.5%) between models. '+A' indicates an addition of the anaphora phenomenon.

| Sports | Conjunction | Comparison | Cause | Relative |
|---|---|---|---|---|
| BERT | 47.0 | 52.3 | 33.0 | 90.9 |
| AugBERT | 46.4 | 74.5 | 41.6 | 90.7 |
| AuxBERT | 47.4 | 53.5 | 33.7 | **91.9** |
| AugAuxBERT | **47.7** | **74.7** | **43.6** | **91.9** |

| Wikipedia | Conjunction | Comparison | Cause | Relative |
|---|---|---|---|---|
| BERT | 55.0 | 67.1 | 33.3 | **91.1** |
| AugBERT | 55.2 | 75.1 | 43.6 | 89.6 |
| AuxBERT | **56.7** | 67.7 | 39.6 | 90.9 |
| AugAuxBERT | 56.6 | **76.0** | **46.0** | 89.4 |

Table 12: Model performance across semantic clusters, measured with MR-EXACT.

not surprising since SR evaluation uses one arbitrary reference, while the augmented dataset guides the model towards balanced fusion variants. Our premise in this paper is that multi-reference evaluation is more adequate in assessing outcomes that paraphrase the original DISCOFUSE fusions. Indeed, the results in Tables 7 and 8 show that with MR evaluation all our models outperform all baselines across setups, with AugAuxBERT, which combines multi-reference training and semantic guidance using auxiliary tasks, performing best.

We further analyze in Table 11 the in-domain model performance of the strongest baseline BERT and our strongest model AugAuxBERT using MR-EXACT, sliced by the different discourse phenomena annotated in DISCOFUSE. As discussed in §4.1, we distinguish between two fusion phenomena types. *Entity-centric* fusion phenomena bridge between two mentions of the same entity, and for such, no connective discourse marker should be added by the model. Our analysis shows that both models perform well on 3 of the 5 *entity-centric* phenomena (bottom part of Table 11). For None and Anaphora, there is a drop in AugAuxBERT performance, which may be attributed to the change in example distribution introduced by our augmented dataset, and will be addressed in future work.

The *semantic relationship* phenomena, on the other hand, require deeper understanding of the relationship between the two input sentences. They tend to be more difficult as they involve the choice of the right connective according to this relation. On these phenomena (top part of Table 11), AugAuxBERT provides substantial improvements compared to BERT, indicating the effectiveness of guiding a model towards a robust semantic interpretation of the fusion task via multiple references and multitasking. Specifically, in the most difficult phenomenon for BERT, discourse connectives, performance increased relatively by 62% for Sports and 26% for Wikipedia. The gap is even larger for the composite cases of discourse connectives combined with anaphora ("Discourse connective+A"): 328.3% (Sports) and 65.4% (Wikipedia).

Finally, to explore the relative importance of the different components of our model, we looked at model performance sliced by the clusters we introduced in §3.2 (see Table 3). The results (Table 12), show that AuxBERT outperforms BERT in 7 of 8 cases, but the gap is $\leq$ 2% in all cases but one. On the other hand, AugBERT improves over BERT mostly for '*Comparison*' and '*Cause*', but the average improvements on these clusters are large: 15.4% (Sports) and 9.2% (Wikipedia). This shows that while our auxiliary tasks offer a consistent performance boost, the inclusion of multiple references contribute to significant changes in model's semantic perception.

## 7 Ablation Analysis

In this analysis, we focus on potential alternative evaluation measures. As mentioned in §2, a possible direction for solving issues in evaluation of sentence fusion—stemming from having a single reference—could be to use similarity-based evaluation metrics (Sellam et al., 2020; Kusner et al., 2015; Clark et al., 2019; Zhang et al., 2020). We notice two limitations in applying such metrics for sentence fusion. First, similarity-based met-

| | Fusion | BERTScore | MR-EXACT |
|---|---|---|---|
| (R) | Ruby is the traditional birthstone for July and is usually more pink than garnet**,** **however** some rhodolite garnets have a similar pinkish hue to most rubies . | - | - |
| (G) | **Although** ruby is the traditional birthstone for July and is usually more pink , than garnet some rhodolite garnets have a similar pinkish hue to most rubies. | 0.9670 | **1** |
| (B) | Ruby is the traditional birthstone for July and is usually more pink than garnet **, thus** some rhodolite garnets have a similar pinkish hue to most rubies . | **0.9893** | 0 |
| (R) | The water level in the wells has risen. **As a result,** work on agricultural lands is going on. | - | - |
| (G) | The water level in the wells has risen**, hence** work on agricultural lands is going on. | 0.9713 | **1** |
| (B) | The water level in the wells has risen. **However,** work on agricultural lands is going on. | **0.9745** | 0 |
| (R) | August 28, **which is the** second day after school starts, is their first away game. | - | - |
| (G) | august 28, **the** second day after school starts, is their first away game. | 0.9834 | **1** |
| (B) | August 28, **who is the** second day after school starts, is their first away game. | **0.9879** | 0 |

Table 13: A demonstration of BERTScore (Zhang et al., 2020) and MR-EXACT (ours) evaluation scores for sentence fusion examples. We mark the ground-truth reference fusion with (R), a correct variant with (G) and an incorrect variant with (B).

rics depend on trainable models that are often in use within fusion models. Thus, we expect these metrics to struggle in evaluation when fusion models struggle in prediction. Second, these metrics fail to correctly account for the semantics of discourse markers, which is particularly important for sentence fusion.

In Table 13 we illustrate typical evaluation mistakes made by BERTScore (Zhang et al., 2020), a recent similarity-based evaluation measure. We calculate BERTScore (F1) for each reference fusion with two variants; (1) a fusion that holds the same meaning and (2) a fusion with a different meaning. A valid evaluation measure for the task is supposed to favor the first option (i.e. the fusion with the same meaning). However, that is not the case for the given examples. The measure often fails to consider the semantic differences between sentences, which is an important element of the task.

Consider the second example in Table 13: BERTScore favours the structural similarity between the gold reference (R) and the incorrect variant (B), which differ in meaning (yet based around the same fusion phenomenon: discourse connective). Meanwhile, the correct variant (G) holds the same meaning as the reference (while a different fusion phenomenon is being used: sentence coordination instead of discourse connective).

## 8 Conclusions

We studied the task of sentence fusion and argued that a major limitation of common training and evaluation methods is their reliance on a single reference ground-truth. To address this, we presented a method that automatically expands ground-truth fusions into multiple references via curated equivalence classes of connective terms. We applied our method to a leading resource for the task.

We then introduced a model that utilizes multiple references by training on each reference as a separate example while learning a common semantic representation for surface form variances using auxiliary tasks for semantic relationship prediction in a multitasking framework. Our model achieves state-of-the-art performance across a variety of evaluation scenarios.

Our approach for evaluating and training with generated multiple references is complementary to an approach that uses a similarity measure to match between similar texts. In future work, we plan to study the combination of the two approaches.

## Acknowledgments

We would like to thank the members of the IE@Technion NLP group and Roee Aharoni, for their valuable feedback and advice. Roi Reichart was partially funded by ISF personal grant No. 1625/18.

## References

Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190, Columbus, Ohio. Association for Computational Linguistics.

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

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 203–213. The Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Elhadad and Kathleen R. McKeown. 1990. Generating connectives. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 97–101.

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.

Katja Filippova and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 177–185. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. Discofuse: A large-scale dataset for discourse-based sentence fusion. In *NAACL-HLT (1)*, pages 3443–3455. Association for Computational Linguistics.

Brigitte Grote and Manfred Stede. 1998. Discourse marker choice in sentence planning. In *Natural Language Generation*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yacine Jernite, Samuel R. Bowman, and David A. Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *CoRR*, abs/1705.00557.

Hongyan Jing and Kathleen R McKeown. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 178–185. Association for Computational Linguistics.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. *CoRR*, abs/1910.00203.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5053–5064, Hong Kong, China. Association for Computational Linguistics.

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, May 7-12, 2018*. European Language Resources Association (ELRA).

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *LREC*. European Language Resources Association.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1006–1017. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Idan Szpektor, Deborah Cohen, Gal Elidan, Michael Fink, Avinatan Hassidim, Orgad Keller, Sayali Kulkarni, Eran Ofek, Sagie Pudinsky, Asaf Revach, et al. 2020. Dynamic composition for conversational domain exploration. In *Proceedings of The Web Conference 2020*, pages 872–883.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

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.

Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In *Proceedings of the Fourth Conference on*

*Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 521–525, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 139–146. Association for Computational Linguistics.

## A Augmentation Rules

In this section we provide the technical details of the augmentation rules used to augment DISCOFUSE (see §3.2). For the sake of clarity, we only provide a general explanation of most rules, avoiding fine-grained issues, minor implementation details and repeating similar rules with minor differences. We note that our augmented dataset will be made publicly available upon acceptance of the paper.

Given a triplet ($s_i^1$, $s_i^2$, $t_i$), where $s_i^1$ and $s_i^2$ are the input sentences and $t_i$ is the ground truth fusion, and its corresponding discourse phenomenon and marker, $p_i$ and $c_i$, respectively, we consider the semantic relationship in $t_i$ which is expressed by $c_i$ (see beginning of table 18). Our augmentation rules relate to three semantic classes: Conjunction, Comparison and Cause, and to one syntactic class: Relative clause (see class definition, §3.2). We design a set of rules for each of these classes, such that each rule first specifies how to detect a fusion that can be augmented according to the rule, and then describes which operations to perform on the ground-truth fusion and its inputs, $t_i$, $s_i^1$ and $s_i^2$, in order to generate a new valid fusion.

We use a set of dictionaries, depicted in Table 14, and a list of pre-defined operations, depicted in Table 15. In Table 16 we provide the technicalities of each rule, presenting its detection and augmentation schema, which is accompanied by the notations and definitions provided in Table 17.

In Table 18 we provide a detailed example of the augmentation process. We start with a description of the input structure, which is detected as a fit for an augmentation rule. We then demonstrate how the variant generation is performed, in a step by step manner.

| Dict | Key |
|---|---|
| $\mathcal{C}_a$ | "furthermore" , "moreover" , "additionally" , |
| $\mathcal{C}_{,a}$ | ", and" |
| $\mathcal{C}_{a,}$ | "furthermore," , "plus," , "additionally," , "moreover," |
| $\mathcal{C}_q$ | "however" , "yet" , "but", "nevertheless" , "although" |
| $\mathcal{C}_{,q}$ | ", yet" , ", but", "although" |
| $\mathcal{C}_{q,}$ | "however," , "still,", "although," "nevertheless," |
| $\mathcal{C}_e$ | "hence" , "therefore" , "consequently" |
| $\mathcal{C}_{e,}$ | "as a result," , "hence," , "thus," , "consequently," , "therefore," |
| $\mathcal{E}_r$ | "who is a", "who is not a", "who is an", "who are an", "who are a", "who is the", "who is not an", "who are not a", "who are the", "which is a", "which is an", "who are not an", "who is not the", "who are not the", "which is not a", "which are an", "which are a", "which is the" "which is not an", "which are not a", "which are not an", "which are the", "which is not the", "which are not the" |
| $\mathcal{P}_r$ | "who is", "who are", "which is", "which are" |

Table 14: The dictionaries used in the data augmentation process.

| Operation | Description |
|---|---|
| Replace($T$, $s_0$, $s_1$) | Replace occurrences of $s_0$ in $T$ with $s_1$. |
| Concat($s_0$, $s_1$) | Attach the string $s_1$ to the end of $s_0$. |
| Delete($T$, $s_0$) | Delete occurrences of $s_0$ from $T$. |

Table 15: Operations on sentences and text phrases, applied for data augmentation (the actual rules are in Table 16). $T$, $s_0$ and $s_1$ are strings, where $T$ is often an entire sentence while $s_0$ and $s_1$ are phrases.

## B Augmentation Statistics

We provide statistics for the entries in our augmented dataset. Table 19 and Table 20 show the distributions of the augmented discourse markers and phenomena, respectively. We note that we have generated a total of 6.5M and 14.7M new fusion examples out of the balanced DISCOFUSE dataset in the Wikipedia and Sports domains, respectively. We then sampled 5M examples of each domain

| Semantic | Detection | Augmentation |
|---|---|---|
| Conjunction | $c_i, c' \in \hat{\mathcal{C}_a}$ | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \in \mathcal{C}_a \wedge c' \in \mathcal{C}_{a,} \wedge$ Concat($.$, $c_i$) $\in t_i$ | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \in \mathcal{C}_a \wedge c' \in \mathcal{C}_{,a} \wedge$ Concat($.$, $c_i$) $\in t_i \wedge len(t) < 40$ | Replace($t_i$, Concat(“.”, $c_i$), $c'$) |
| Comparison | $c_i, c' \in \hat{\mathcal{C}_q}$ | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \notin \mathcal{C}_q, \wedge c' \in \mathcal{C}_q, \wedge p =$ S-coordination | Replace($t_i$, $c_i$, Concat(“.”, $c'$)) |
| | $c_i \notin \mathcal{C}_{,q} \wedge c' \in \mathcal{C}_{,q} \wedge p =$ Inner-connective $\wedge c' \in$ {but, yet} | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \notin \mathcal{C}_q, \wedge c' \in \mathcal{C}_q, \wedge p =$ Inner-connective | Replace($t_i$, $c_i$, Concat(“.”, $c'$)) |
| | $c_i \notin \mathcal{C}_q, \wedge c' \in \mathcal{C}_q, \wedge p =$ Discourse-connective | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \notin \mathcal{C}_{,q} \wedge c' =$ {, although} $\wedge p =$ Discourse-connective | Replace($t_i$, Concat(“.”,$c_i$), $c'$) |
| | $c' \in \mathcal{C}_{,q} \wedge c_i =$ {although} $\wedge p =$ Forward-connective | Concat(Concat(Delete($s_i^1$, “.”), $c'$), $s_i^2$) |
| Cause | $c_i, c' \in \hat{\mathcal{C}_e}$ | Replace($t_i$, $c_i$, $c'$) |
| | $c_i \in \mathcal{C}_e \wedge c' \in \mathcal{C}_{e,} \wedge$ Concat($.$, ) $\in t_i \wedge$ Concat($.$, $c_i$) $\notin t_i$ | Replace(Delete($t_i$, $c_i$), “.”, Concat(“.”,$c'$)) |
| Relative Clause | $p =$ Relative Clause $\wedge \exists a \in \mathcal{E}_r \mid a \in t_i$ | Delete($t_i$, $b$), $b \in a \cap \mathcal{P}_r$ |

Table 16: Augmentation rules for derivations of new fusions out of DISCOFUSE ground-truth fusions. We mark with red the rule discussed in the detailed augmentation example in Table 18.

| Notation | Definition |
|---|---|
| $t_i$ | The ground-truth fusion of the $i$-th example |
| $s_i^1$, $s_i^2$ | The two input sentences of the $i$-th example |
| $c_i$ | The discourse marker used in $t_i$ |
| $p_i$ | The discourse phenomenon of $t_i$ |
| $\mathcal{C}_a$ | A list of conjunction markers, without a comma |
| $\mathcal{C}_{a,}$ | A list of conjunction markers with a right comma |
| $\mathcal{C}_{,a}$ | A list of conjunction markers with a left comma |
| $\mathcal{C}_q$ | A list of comparison markers, without a comma |
| $\mathcal{C}_{q,}$ | A list of comparison markers with a right comma |
| $\mathcal{C}_{,q}$ | A list of comparison markers with a left comma |
| $\mathcal{C}_e$ | A list of cause and effect markers, without a comma |
| $\mathcal{C}_{e,}$ | A list of cause and effect markers with a right comma |
| $\mathcal{E}_r$ | A set of relative clause expressions which can transform to an apposition phrases without adding any tokens |
| $\mathcal{P}_r$ | A set of relative clause pronouns |

Table 17: Notations and definitions for Table 16.

for training AugBERT and AugAuxBERT. These tables provide details about the imbalanced augmentation, where specific phenomena and markers are generated more often than others during the augmentation process.

## C Probability Distribution across Valid Fusions

According to the results our models achieve in MR evaluation, we conclude that they are better capable of generating a fused text that is included in the ground-truth set. Here we show that, in addition, they assign a more uniform probability to the members of the set, compared to the BERT model. Figure 2 graphically illustrates this pattern for three typical examples with 9, 9 and 5 ground-truth fusions, respectively (in each example, fusion 1 is the one in the original DISCOFUSE, and the others were created in our expansion).

We first formally show that the probability mass tends to be uniformly allocated among the various references; for any $t \in \mathcal{V}(t_i)$ let

$$\bar{p}_i(t) = \frac{p(t|s_i^1, s_i^2)}{\sum_{t' \in \mathcal{V}(t_i)} p(t'|s_i^1, s_i^2)}$$

be the probability of a variant $t$ relative to the overall probabilities of the variants in $\mathcal{V}(t)$. Indeed, for more than 99% of the test-set examples the entropy $-\sum_{t \in \mathcal{V}(t_i)} \bar{p}_i(t) \log \bar{p}_i(t)$ induced by AugBERT and AugAuxBERT for the ground-truth solutions is higher than that of BERT, indicating that our augmented models are less inclined to prefer one of the solutions over the others.

Moreover, we computed the sum of the multiple-reference probabilities $\sum_{t \in \mathcal{V}(t_i)} p(t|s_i^1, s_i^2)$ in test-set examples. In about 71% of the test-set examples the sum of probabilities induced by AugBERT and AugAuxBERT is higher than that of BERT. That is, our model learns to direct more overall probability mass towards the correct variants, indicating a higher confidence in the correct solutions.

## D Hyper-Parameters and Configurations

The BERT, AuxBERT, AugBERT and AugAuxBERT models share the same hyper-parameters with respect to their shared architecture and to the optimization process. All models use

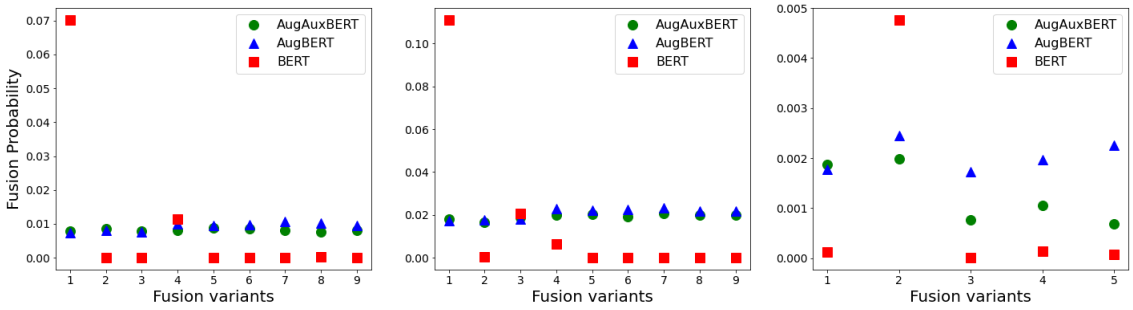


Figure 2: The probability assigned to the ground-truth fusions by our AugBERT and AugAuxBERT models, and by the baseline BERT, for three typical DISCOFUSE examples. Our models assign higher and more uniform probabilities to the members of the ground-truth set.

**1. Input: Ground-truth fusion**

$s_i^1$ = {The company had bigger facilities at Wembley in the west of the capital.}
$s_i^2$ = {It was easier to attract stars and audiences to central London.}
$t_i$ = {Although the company had bigger facilities at Wembley in the west of the capital, it was easier to attract stars and audiences to central London.}
$c_i$ = although
$p_i$ = Forward-connective

**2. Detection**

$c' = \{, \text{but}\} \in \mathcal{C}_{,q} \wedge c_i = \{\text{although}\} \wedge p_i =$ Forward-connective

**3. Operations**

| | |
|---|---|
| DELETE$(s_i^1, \text{“.”})$ | $X_1$ = {The company had bigger facilities at Wembley in the west of the capital} |
| CONCAT$(X_1, c')$ | $X_2$ = {The company had bigger facilities at Wembley in the west of the capital, but} |
| CONCAT$(X_2, s_i^2)$ | $X_3$ = {The company had bigger facilities at Wembley in the west of the capital, but it was easier to attract stars and audiences to central London.} |

**4. Output - augmented fusion**

$t_i'$ = {The company had bigger facilities at Wembley in the west of the capital, but it was easier to attract stars and audiences to central London.}
$c_i'$ = , but
$p_i'$ = Sentence-coordination

Table 18: A detailed augmentation rule execution example. We mark discourse markers in red. The ground-truth fusion $t_i$ consists of the input together with the two source sentences, $s_i^1$ and $s_i^2$.

| Sports | | Wikipedia | |
|---|---|---|---|
| | % | | % |
| although | 18.8 | still | 15.1 |
| yet | 16.7 | although | 24.4 |
| nevertheless | 16.0 | nevertheless | 15.9 |
| still | 14.0 | however | 10.6 |
| however | 13.1 | yet | 15.6 |
| but | 8.4 | but | 8.4 |
| consequently | 1.1 | hence | 1.3 |
| moreover | 1 | consequently | 1.2 |

Table 19: The most common connectives augmented to the balanced DISCOFUSE dataset. Percentages are calculated with respect to the entire set of new fusions in each domain.

| Discourse phenomena | Sports (%) | Wiki(%) |
|---|---|---|
| VP coordination | 7.8 | 6.8 |
| Inner connective | 9.5 | 6.1 |
| Inner connective+A | 2.4 | 2.7 |
| Sentence coordination | 12.1 | 11.2 |
| Sentence coordination+A | 3.1 | 4.3 |
| Discourse connective | 48.9 | 44.8 |
| Discourse connective+A | 15.7 | 23.7 |
| Apposition | 0.2 | 0.1 |

Table 20: Discourse phenomena distribution of augmented fusions from the balanced DISCOFUSE dataset. '+A' indicates an addition (composition) of the anaphora phenomenon, and Wiki stands for Wikipedia.

an initialized BERT-Base Uncased encoder with a randomly initialized Transformer (Vaswani et al., 2017) decoder. Configuration details and the hyper-parameters of the training process are provided in Table 21.

Recall that we define our multi-task loss as follows:

$$\ell_{\text{total}} = \ell_{\text{gen}} + \alpha \cdot \ell_{\text{type}} + \beta \cdot \ell_{\text{conn}}$$

where $\ell_{\text{gen}}$ is the cross-entropy loss of the genera-

| Encoder - BERT | |
|---|---|
| hidden size | 768 |
| number of attention heads | 12 |
| number of hidden layers | 12 |
| vocab size | 30522 |
| hidden activation | *'gelu'* |
| number of parameters | 108891648 |
| **Decoder - Transformer** | |
| hidden size | 768 |
| number of attention heads | 8 |
| number of hidden layers | 6 |
| number of parameters | 47238144 |
| **Classifiers - Pooler** | |
| input dim | 768 |
| first hidden dim | 768 |
| second hidden dim | 256 |
| phenomena output dim | 13 |
| connective output dim | 71 |
| number of parameters | 809044 |
| **Optimization** | |
| optimizer | *AdamOptimizer* |
| beta1 | 0.9 |
| beta2 | 0.997 |
| epsilon | $1e-9$ |
| batch size | 100 |

Table 21: The hyper-parameters of the BERT, AuxBERT, AugBERT and AugAuxBERT models.

| | W | S |
|---|---|---|
| BERT | 63.5 | 60.4 |
| AuxBERT | **64.3** | **61.4** |
| AugBERT | 52.2 | 51.0 |
| AugAuxBERT | 52.8 | 52.3 |

Table 22: Single reference EXACT results on development data.

tion task, and $\ell_{\text{type}}$ and $\ell_{\text{conn}}$ are the cross-entropy losses of the discourse type and connective phrase predictions, respectively, with scalar weights $\alpha$ and $\beta$. We tuned $\alpha$ and $\beta$ on the DISCOFUSE development sets, considering the values $\{0.1, 0.5, 1\}$ for both weights. We then chose the best performing set of hyperparameter according to the higher EXACT score on the appropriate development set. In all cases the resulting values were 0.1 for both weights.

The auxiliary heads of AuxBERT and AugAuxBERT also share the same architecture and hyper-parameters. For each auxiliary classifier we used one fully-connected layer, where the input dimension is 768, derived from *BERT's pooler* output, and the output dimension is determined by the auxiliary output dimension (71 discourse markers and 12 discourse phenomena).

We use the best-performing architecture and hyper-parameters specified by Malmi et al. (2019) for the LaserTagger model. Specifically, we use the auto-regressive model, AR-LaserTagger, with an initialized BERT-Base Cased encoder and a small randomly initialized Transformer decoder. This model has shown better results on the fusion task compared to FF-LaserTagger, the non auto-regressive model.

## E Experimental Details

All experiments were performed on either one or two Nvidia GeForce GTX 1080 Ti GPUs, with two cores, 11 GB GPU memory per core, 6 CPU cores and 62.7 GB RAM.

We measured an average of 8.5 hours for 45,000 training steps for BERT, AuxBERT and AugAuxBERT, which is approximately a full 'Wikipedia' epoch and about one-third of a full 'Sports' epoch. To achieve full convergence, each model requires about 675K-900K training steps.

In Table 22 we provide the corresponding single-reference EXACT validation performance for each reported test result. Notice that domain adaptation setups are not included within this table, since in such setups we use development data from the source domain.

## F URLs of Code and Data

As noted in §5, we provide here the URLs for the datasets and code we have used:

- DISCOFUSE (Geva et al., 2019) A large scale dataset for sentence fusion: https://github.com/google-research-datasets/discofuse
- Code and pre-trained weights of the pre-trained *BERT-Base, Uncased* (Devlin et al., 2019) model: https://github.com/google-research/bert
- Code for LaserTagger (Malmi et al., 2019): https://github.com/google-research/lasertagger
- Code for BERTScore (Zhang et al., 2020): https://github.com/huggingface/nlp/metrics/bertscore