applied sciences

MDPI

Article

Automatic Construction of Fine-Grained Paraphrase Corpora System Using Language Inference Model

Ying Zhou †, Xiaokang Hu † and Vera Chung *

School of Computer Science, Faculty of Engineering, Darlington Campus, The University of Sydney, Darlington, NSW 2008, Australia; ying.zhou@sydney.edu.au (Y.Z.); xihu8986@uni.sydney.edu.au (X.H.)
* Correspondence: vera.chung@sydney.edu.au
† These authors contributed equally to this work.

Abstract: Paraphrase detection and generation are important natural language processing (NLP) tasks. Yet the term paraphrase is broad enough to include many fine-grained relations. This leads to different tolerance levels of semantic divergence in the positive paraphrase class among publicly available paraphrase datasets. Such variation can affect the generalisability of paraphrase classification models. It may also impact the predictability of paraphrase generation models. This paper presents a new model which can use few corpora of fine-grained paraphrase relations to construct automatically using language inference models. The fine-grained sentence level paraphrase relations are defined based on word and phrase level counterparts. We demonstrate that the fine-grained labels from our proposed system can make it possible to generate paraphrases at desirable semantic level. The new labels could also contribute to general sentence embedding techniques.

Keywords: paraphrase; text generation; language inference

Citation: Zhou, Y.; Hu, X.; Chung, V. Automatic Construction of Fine-Grained Paraphrase Corpora System Using Language Inference Model. *Appl. Sci.* **2022**, *12*, 499. https://doi.org/10.3390/app12010499

Academic Editor: Valentino Santucci

Received: 21 November 2021
Accepted: 28 December 2021
Published: 5 January 2022

1. Introduction

Paraphrase detection and generation are important natural language processing (NLP) tasks. There are a few widely used benchmark datasets constructed through automatic extraction and manual labelling. The accuracy of a manual label is domain and task dependent. In some cases, raters are asked to judge if two sentences are "semantically equivalent". The choice of label depends on a rater's tolerance of semantic divergence. As stated in [1], many sentence pairs judged as "semantically equivalent" diverge semantically to some degree. Paraphrase relation, by definition, is a symmetric bidirectional entailment relation. With the presence of semantic divergence, the relation becomes a directional forward or reverse entailment.

Table 1 shows different directional sentence relations from Microsoft Research Paraphrase Corpus (MRPC) which are all labelled as paraphrase:

Table 1. Different sentence relations with positive paraphrase label from MRPC.

	Sentence 1	Sentence 2
1	Amrozi accused his brother , whom he called "the witness", of deliberately distorting his evidence.	Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
2	PeopleSoft also said its board had officially rejected Oracle's offer.	Thursday morning, PeopleSoft's board rejected the Oracle takeover offer.

In the first sentence pair, the amount of information in the two sentences are equivalent. Using either sentence as the premise, we can derive that the other is true. This is sometimes referred to as bidirectional entailment. The second pair is an example of reverse entailment

where sentence 1 can be derived from sentence 2 but not the other way round because sentence 2 contains more information than sentence 1.

A dataset with a more strict rule may label the second sentence pair as a negative case. Such variation can affect the generalisability of a paraphrase classification model. The presence of both symmetric and directional relations in a single class also affects the predictability of a paraphrase generation task. A generation model trained with randomly mixed relations would generated results of random relations.

This paper proposes a novel method to automatically generate fine-grained paraphrase labels using language inference models. In particular, we make the following contributions:

- We defined a set of fine-grained sentence-level paraphrase relations based on similar relations at the word and phrase level.
- We developed a method utilising the language inference model to automatically assign fine-grained labels to sentence pairs in existing paraphrase and language inference corpora.
- We demonstrated that models trained with fine-grained data are able to generate paraphrases with specified directions.

The labelling process leads to a detailed examination of several public corpora. We discover that corpora constructed for similar linguistic tasks have very different compositions of fine-grained relations. For instance, we find that:

- Compared with Quora Question Pair (QQP), MRPC tolerates more semantic divergence in its positive class, which contains more directional paraphrases than equivalent ones.
- Compared with the Stanford Natural Language Inference (SNLI), Multi-Genre Natural Language Inference (MNLI) contains more diversified sentence pairs in all three classes.

Such information may help researchers to design customised optimisation and to provide insights on observed performance variation.

2. Related Work

Although the concept of paraphrase has been around for a long time, there has been no precise and widely accepted definition of paraphrase. The multiple definitions found in different literature can be viewed as "paraphrases" of each other. It is defined as "approximate conceptual equivalence among outwardly different materials" in an early linguistics text *Introduction to Text Linguistic* [2]. More recent literature from computational linguistics provided definitions such as "expressing one thing in other words" [3] , or "alternative ways to convey the same information" [4].

Basically, paraphrases are sentences which convey the same or similar meanings. Although paraphrases, in a strict and narrow view, require completely semantically equivalence, most researchers take a broader view of paraphrase, allowing more flexibility and approximate equivalence known as 'quasi-paraphrase' [5]. The example from the [6] study shows that sentences 1 and 2 are considered as paraphrases even though they contain a slightly different amount of information. This, however, blurs the boundary between paraphrase and non-paraphrase, making a paraphrase processing system hard to build.

1. Authorities said a young man injured Richard Miller.
2. Richard Miller was hurt by a young man.

The study of paraphrase can be useful for many natural language processing applications. In text summarisation, the information repeated across multiple documents is extracted through the identification of paraphrase sentences [7]. Paraphrase identification can also be used in plagiarism detection. In natural language generation, the coherence and fluency improve when having more varied paraphrased candidate sentences. For example, paraphrase can be used to convert the professional terminology to simple text so that non-experts can understand [8]. It also allows questions with similar meaning to be expressed differently, which could improve system efficiency considerably. For example, in

a question-answering system, a random user input question could be mapped to some of the frequently asked questions (FAQs) with the help of a paraphrase identification tool [9]. In addition, paraphrase identification can be used to recognise duplicate questions for online question and answer (QA) forums to combine and redirect similar questions.

Other semantic relationships exists between sentences, among which, the entailment relation is closely related to paraphrase. Given two sentences, if a hypothesis sentence can be inferred from the given premise, we can say that the premise sentence entails the hypothesis sentence. For instance, sentence 1 entails sentence 2 in the example below.

1. **Premise**: A soccer game with multiple males playing.
2. **Hypothesis**: Some men are playing a sport.

From the entailment relation perspective, paraphrase can be viewed as a special case of the entailment relationship where two sentences are bidirectionally entailed [10]. In other words, sentence A is a paraphrase of sentence B only if A entails B and B entails A. This perspective provides a solution for the paraphrase identification. However, in most cases, a paraphrase, as discussed in the previous section, is usually a quasi-paraphrase with a certain content loss. Limiting paraphrase to bidirectional entailment reduces a large proportion of cases [11].

Several studies had been aimed at extracting paraphrase data [3,4]. Microsoft Research Paraphrase Corpus (MRPC) is paraphrase corpus which is the most widely used benchmarks for paraphrase identification. It contains 5801 sentence pairs extracted from online news articles [1]. The benchmark itself only has one simple binary label indicating whether the pair is a paraphrase. The judgement depends on the rater's tolerance on the semantic divergence. The question-answer community `Quora` released Quora Question Pairs [12], which contains more than 400,000 question pairs. The question pairs sharing the same semantics are annotated as a duplicate based on the judgement of users and Quora's merging policy. Another dataset, the semantic textual similarity benchmark (STS-B), alleviates the problem by a human-annotated semantic similarity score ranged from 1 to 5 [13]. This dataset still needs human judges with a specific linguistic proficiency.

There has been work showing that sentence pairs can potentially have different varieties of paraphrase relations [14]. Bhagat et al. [15] originally point out that the paraphrase inference rule is underspecified in directionality. The researchers define three plausible inference rules and implement the algorithm least-disruptive topology repair (LEDIR) to classify the directionality of inference rules. With the directionality hypothesis that paraphrase statements tend to appear in similar contexts, LEDIR measures the context similarity of statements. The Paraphrase Database (PPDB) 2.0 also shows that paraphrase word level pair can have explicit entailment relationships. Pavlick et al. [16] annotate paraphrase pairs with an explicit entailment relation based on natural logic. However, sentence-level paraphrase relationship needs further research.

3. Auto Relabelling Methods

3.1. Fine-Grained Paraphrase Relations

Fine-grained paraphrase relations have been investigated at the word and phrase level. PPDB, a large paraphrase dataset at the word and phrase level, introduced fine-grained entailment relation in version 2.0 [14]. The seven basic entailment relationship used in PPDB 2.0 were formally defined in Bill MacCartney's thesis "Natural language inference" [17]. These include: equivalence ($\equiv$), forward entailment ($\sqsubset$), reverse entailment ($\sqsupset$), negation ($\wedge$), alternation ($|$), cover ($\smile$) and independence (#). MacCartney tried to map various entailment relations to the 16 elementary set relations. Among the 16 elementary set relations, 9 involve an empty set or universe and are discarded in the language entailment domain, which leaves 7 meaningful relations. Examples of each relation as taken from MacCartney's thesis are reproduced in Table 2.

Table 2. Word level entailment relation examples in MacCartney's thesis [17].

Symbol	Name	Example
$x \sqsubset y$	forward entailment	$crow \sqsubset bird$
$x \sqsupset y$	reverse entailment.	$Asian \sqsupset Thai$
$x \equiv y$	equivalence	$couch \equiv sofa$
$x \mid y$	alternation	$cat \mid dog$
$x \wedge y$	negation	$able \wedge unable$
$x \smile y$	cover	$animal \smile non\text{-}ape$
$x \# y$	independence	$hungry \# hippo$

Most of those relations also exist at the sentence level. A positive paraphrase class typically includes four of the above relations: equivalence, forward entailment, reverse entailment and alternate. A negative paraphrase class may contain all except the equivalent relation.

Table 3 shows the four relations with example sentence pairs from the STS-B dataset. All pairs have a score above 4.5, indicating strong semantic similarity. The equivalence relation is sometimes referred to as bidirectional entailment. It indicates true paraphrase, in which the two sentences contain the same amount of information but expressed in different ways. As we seldom have words with exactly the same meaning, the paraphrase may still have small semantic divergence at the word level. The alternate relation example shows two sentences share the key message but each with its own piece of details. This satisfies the set theory definition of alternate relation as two sets with non-empty intersection and their union is not the universe. The alternate relation is general enough to include sentence pairs with less overlapping semantics to be considered as paraphrase. We focus on the first three well-defined relations.

Table 3. Fine-grained paraphrase example.

Relation	First	Second
$\sqsubset$	A young child is riding a horse.	A child is riding a horse.
$\sqsupset$	Three men are playing guitars.	Three men are on stage playing guitars.
$\equiv$	The man cut down a tree with an axe.	A man chops down a tree with an axe.
$\mid$	The report also claims that there will be up to 9.3 million visitors to hot spots this year, up again from the meagre 2.5 million in 2002.	There will be 9.3 million visitors to hot spots in 2003, up from 2.5 million in 2002, Gartner said.

3.2. Observations from Language Inference Datasets

Instead of manually labelling, we propose a method utilising the existing labels in language inference dataset. In particular, we utilise datasets with three labels: "entailment", "neutral" and "contradiction". There are two such datasets: the Stanford Natural Language Inference corpus (**SNLI**) [18] and the Multi-Genre Natural Language Inference (**MNLI**) corpus [19]. They are constructed following a similar process where AMTs are given a prompt sentence (the premise) and are asked to write three sentences (the hypotheses): one that is definitely true ("entailment"; one that is probably true ("neutral") and one that is definitely not true ("contradiction"). The SNLI corpus contains 570,000 sentence pairs sourced from image captions. The MNLI corpus [19] is of an approximately similar size but with various genres of spoken and written text, including transcripts, government reports and fictions. The construction process ensures that both datasets maintain balanced samples in the three classes.

The three classes only represent a subset of all possible sentence relations. Each class may contain sentence pairs belonging to a few fine-grained relations. In particular, we observe that:

- The "entailment" label contains forward entailment ($\sqsubset$) and equivalent ($\equiv$) pairs. This conforms to the definition of "entailment" class. The first two rows in Table 4 show sample "entailment" pairs from MNLI. The first pair is an example of forward entailment while the second one is an example of bidirectional entailment.
- The "neutral" class contains reverse entailment pairs, alternate pairs, independent pairs and pairs of other relations. It seems to be the result that "neutral" is designed as a catch-all class. There are also cases where event and entity coreference could make the distinction between "neutral" and "contradiction" ambiguous. The last three rows in Table 4 shows three sample "neutral" pairs from MNLI.

Table 4. Sample "entailment" and "neutral" pairs from MNLI.

Relation	Premise	Hypothesis
$\sqsubset$	Finally, the FDA will conduct workshops, issue guidance manuals and videotapes and hold teleconferences to aid small entities in complying with the rule.	The FDA is set to conduct workshops.
$\equiv$	Postal Service were to reduce delivery frequency.	The postal service could deliver less frequently.
$\sqsupset$	A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.
#	He went down on his knees, examining it minutely, even going so far as to smell it.	It smelled like eggs.
|	The company once assembled, Poirot rose from his seat with the air of a popular lecturer, and bowed politely to his audience.	Poirot rose from his seat, bowed and started addressing the audience.

Such observations enable us to define rules for extracting fine-grained sentence pair relations from existing datasets.

3.3. Auto Relabel Rules

Most fine-grained paraphrase relations are symmetric; these include: equivalence, alternation, independence and negation. The forward and reverse entailment are directional. This suggests: If a sentence pair (s_1, s_2) has symmetric relation, the swapped pair (s_2, s_1) should have the same relation; if a sentence pair (s_1, s_2) has forward entailment relationship, the swapped pair (s_2, s_1) should have reverse entailment relation and could be labelled as "neutral".

We utilise the labels of swapped pairs to identify directional and symmetric entailment relations. The rules are defined as follows: for a sentence pair, (s_1, s_2):

1. If (s_1, s_2) is "contradiction" and (s_2, s_1) is "contradiction", (s_1, s_2) is of negation or contradiction relation and is labelled as 0.
2. If (s_1, s_2) is "entailment" and (s_2, s_1) is "neutral", (s_1, s_2) is a $\sqsubset$ relation and is labelled as 1.
3. If (s_1, s_2) is "neutral" and (s_2, s_1) is "entailment", (s_1, s_2) is a $\sqsupset$ relation and is labelled as 2.
4. If (s_1, s_2) is "entailment" and (s_2, s_1) is "entailment", (s_1, s_2) is of equivalent or bidirectional entailment relation ($\equiv$) and is labelled as 3.

5. If (s_1, s_2) is "neutral" and (s_2, s_1) is "neutral", the sentence pair could be of alternation or independence relation. We label it as 4.

The rules cover a few combinations of original and swapped pair labels. Not all combinations are semantically possible. The definition of "entailment" and "contradiction" make it impossible for a sentence pair and swapped pair to have these combinations. However, the fuzzy boundary between "contradiction" and "neutral" caused by entity or event coreference [18,20] may introduce cases of such combinations in the dataset.

It is semantically not correct to have the original pair classified as "contradiction" and the swapped pair classified as any relation other than "contradiction". If some impossible combination appears in the dataset, it could be caused by either the classification error or labelling error. If each dataset contains some pairs with an impossible combination, this could be used to optimise the classifier.

4. Automatic Relabelling with Fine-Grained Paraphrase Relations

We construct the corpora from existing datasets containing sentence pairs with all fine-grained paraphrase relations. These include the language inference datasets as well as paraphrase datasets. After examining the size and sentence relations in various datasets, we focus the construction efforts on two language inference datasets: MNLI and SNLI and three paraphrase datasets: MRPC [1], QQP [12] and STS-B [13].

MRPC The Microsoft Research Paraphrase Corpus [1] consists of 5000+ sentence pairs extracted from Web news. The sentences pairs are manually annotated with a binary label to indicate positive or negative paraphrase relation. The dataset has around 33% negative samples and the others are all positive.

QQP The Quora Question Pairs [12] is a dataset with more than 400,000 question pairs released by the question-answer community `Quora`. The question pairs sharing the same semantics are annotated as duplicate based on Quora's merging policy. Similar to MRPC, the data is unbalanced, in which 63% are negative duplicates.

STS-B Instead of using binary label representing the semantic equivalence of the sentence pairs, the semantic textual similarity benchmark [13] uses a similarity score ranged from 0 to 5. The corpus is extracted from news headlines, video and image captions data. Similar to MRPC and QQP, each pair is human-annotated.

SNLI The Stanford Natural Language Inference (SNLI) corpus ([18]) is a collection of sentence pairs labelled for three classes including entailment, contradiction, and semantic independence. SNLI is large corpus resources which has 570,152 sentence pair. In addition, all of its sentences and labels were written by humans in a grounded, naturalistic context. There are four additional judgements for each label for 56,941 of the examples. Of these, 98% of cases emerge with at least three annotator consensus.

MNLI The Multi-Genre Natural Language Inference (MNLI) corpus ([19]) is 433,000 sentence pairs from a broad range of genres of written and spoken English, annotated with sentence inference classes and balanced across three labels. In the corpus, each premise sentence is derived from one of ten sources of text, which constitute the ten genre sections of the corpus while each hypothesis sentence and pair label was composed by a crowd worker in response to a premise. The corpus is modelled on the SNLI corpus but differs in that it covers a range of genres of spoken and written text and supports a distinctive cross-genre generalisation evaluation.

Table 5 summarises the label types of a different dataset.

Table 5. Datasets label types being used for relabelling.

Data Set	Label Type	Label Examples
MRPC	discrete	[0, 1]
QQP	discrete	[0, 1]
STSB	continuous	[0, 5.0]
SNLI	discrete	[“contradiction”, “neutral”, “entailment”]
MNLI	discrete	[“contradiction”, “neutral”, “entailment”]

4.1. Three-Label Language Inference Classifiers and Initial Data Cleansing

The relabelling practice starts by assigning one of three entailment labels to both the original sentence pairs in the dataset and the corresponding swapped ones. For MNLI and SNLI, the assignment is required only for the swapped ones. For a paraphrase dataset, entailment labels need to be assigned to both the original and swapped pairs. The final fine-grained label is determined by the combination of the two labels following the rules introduced in Section 3.3.

Two classifiers, one trained on MNLI and the other trained on SNLI, are used to assign entailment labels. For MNLI classifier, we use a pre-trained **RoBERTa** [21] model included in PyTorch huggingface project:

https://github.com/huggingface/transformers (accessed on 23 November 2020). We train our own SNLI classifier with RoBERTa. RoBERTa is a variation of BERT which uses the same transformer encoder architecture but with a few modifications of the pre-training process. During pre-training, RoBERTa removes NSP training task, dynamically changes the masking pattern in MLM and pre-trains the model for larger data for a longer time.

The SNLI classifier is trained with the experiment setup in Table 6. We implement a random search from coarse to fine to find the optimal combination of hyperparameters. Afterwards, we choose the optimal combination based on the development dataset performance keeping random seed value fixed.

Table 6. Experimental setup for hyperparameter tuning.

RoBERTa	Roberta-Large
Maximum Input Token	{256}
Learning Rate	$\{1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}\}$
Batch sizes	{8, 16}

Table 7 shows the accuracy comparison of the two classifiers. The automatic construction process only keeps the sentence pairs that the two classifiers agree on both labels for paraphrase datasets. It keeps the sentence pairs that the two classifiers agree on the reversed labels for entailment datasets. The MNLI classifier seems to be more versatile than the SNLI classifier. It performs well on both SNLI and MNLI datasets.

Table 7. Accuracy Comparison of Two Classifiers.

Classifier	SNLI	MNLI Matched	MNLI Mismached
SNLI-RoBERTa	89.65%	77.04%	76.99%
MNLI-RoBERTa	87.55%	89.46%	89.15%

4.2. Summary Statistics of Fine-Grained Labels

Table 8 shows the percentage of various original-swapped pair label combinations in the five datasets. As expected, the semantically impossible combinations of “contradiction”

and "entailment" only occur less than 1% in all datasets. The combinations of "neutral" and "contradiction" do occur quite often in all datasets as the result of an ambiguous boundary between "neutral" and "contradiction". The percentage of such combination is different in different datasets. For instance, the SNLI has only 4.3% of contradiction–neutral pairs, while MNLI has 17% of such a pair. We suspect that a dataset with a relatively narrow language domain, such as image caption, may have less ambiguity in these two labels, leading to a cleaner class membership.

Table 8. Percentage of label combinations in various datasets.

Relation Combination	MRPC	QQP	STSB	SNLI	MNLI
Entailment–Entailment	9.00%	14.00%	9.00%	2.30%	7.10%
Entailment–Neutral	15.50%	10.80%	10.00%	28.70%	18.70%
Entailment–Contradiction	0.30%	0.6%	0.40%	0.30%	0.70%
Neutral–Entailment	16.50%	10.30%	10.60%	3.20%	3.80%
Neutral–Neutral	35.60%	27.40%	8.50%	27.60%	25.20%
Neutral–Contradiction	4.60%	8.30%	5.30%	3.10%	6.40%
Contradiction–Entailment	0.20%	0.50%	0.50%	0.26%	0.70%
Contradiction–Neutral	4.30%	7.70%	5.70%	4.30%	17.00%
Contradiction–Contradiction	14.00%	20.30%	40.00%	30.20%	20.40%

The fine-grained label also reveals that the distribution of sentence relations are very different in different datasets. STS-B has the most balanced directional and bidirectional entailment relation distribution. It also has a large portion of negation relations. Compared with QQP, MRPC seems tolerate more semantic divergence in its positive paraphrase class. It only has 9% of equivalent relation but 15.5% and 16.5% of forward and reverse entailment relation, respectively. While QQP has 14% equivalence relation and only 10.8% and 10.3% of forward and reverse entailment relation. Both have relatively large percentages of alternation and independent relations.

SNLI and MNLI, though constructed following the same practice, also differ in terms of fine-grained sentence relations. SNLI conforms more to the strict definition of each class. Its entailment class contains mostly pairs with forward entailment relation with only a small portion of equivalent relation. The overall percentages are 28.7% and 2.3%, respectively. MNLI's entailment class contains more equivalent relation (7.1%) and less forward entailment relation (18.7%). They each have a small percentage of sentence pairs in "neutral" class that are of reverse entailment relation. We can also see that most of SNLI's original "contradiction" class contains sentence pairs of symmetric negation relation, while MNLI's original "contradiction" class contain a large percentage of pairs of asymmetric relations.

Table 9 shows the sentence pair count of each fine-grained class in the five datasets. We discard all the impossible sentence relations as discussed previously and all the sentence pairs on which the two classifiers disagree. The overall fine-grained data size is proportional to the original data size.

Table 9. Fine-grained class sentence counts.

Fine-Grained Class	MRPC	QQP	STSB	SNLI	MNLI
0 (∧)	197	46,156	1389	117,614	52,185
1 (⊏)	400	29,413	450	127,093	59,126
2 (⊐)	427	27,982	479	14,152	11,942
3 (≡)	233	38,160	406	10,389	22,395
4 (|or#)	846	42,102	688	86,263	63,501

5. Fine-Grained Label Correctness and Accuracy Investigation

In this section, we investigate the correctness of the relabelling practice utilising the three paraphrase datasets' original classes and scores. We cross-check that the property of the newly assigned class conforms to the original class and similar score. We also investigate the string property of each dataset.

5.1. Original Label vs. Fine-Grained Label

Table 10 shows the properties of fine-grained classes with respect to each dataset's original class or value. For STS-B, the average similarity score of each new label is calculated. For MRPC and QQP, the positive paraphrase percentage of each new label is calculated.

Table 10. Positive paraphrase percentage of each fine-grained paraphrase label.

		Fine-Grained Class Label				
Metrics	**Data Sets**	**0 (∧)**	**1 (⊏)**	**2 (⊐)**	**3 (≡)**	**4 (|or#)**
Average Similarity Score	**STS-B**	1.19	3.69	3.71	4.62	3.06
Positive Paraphrase Percent	**MRPC**	39.6%	81%	82.0%	99.1%	49.9%
	QQP	6.4%	52.6%	56.7%	86.2%	20.8%

The results on STS-B are consistent with the definition of new labels. The pairs with equivalence (Label 3) have the highest average similarity score: 4.62. Directional entailment pairs have lower average similarity scores and there is not much difference in similarity score between forward (3.69) and reverse entailment (3.71). Label 4 has a lower average than the other three as the relation is either independence or alternation. Label 0 represents negation relation. It has the lowest similarity score. This suggests that our relabelling rule is able to identify the semantic difference among fine-grained classes.

The results on MRPC and QQP are in general as expected. Both datasets contain pairs belonging to various entailment relations in the positive class. We expect to see a very high percentage of positive paraphrases in the newly assigned equivalent class (≡). We also expect to see a relatively high percentage of positive paraphrases in the other two entailment classes (⊏ and ⊐). In the MRPC dataset, 99.1% of the pairs labelled as ≡ are positive paraphrases, while 81% of the pairs labelled as ⊏ and 82% of the pairs labelled as ⊐ are positive paraphrases. In the QQP dataset, 86.2% of the pairs labelled as ≡ are positive paraphrases, while 52.6% of the pairs labelled as ⊏ and 56.7% of the pairs labelled as ⊐ are positive paraphrases.

Overall, the results confirms that MRPC tolerates more semantic divergence in positive labels. It has a higher percentage of forward and reverse entailment pairs labelled as positive paraphrases. This might be the results of different annotation practices adopted while constructing the dataset.

Theoretically, all pairs labelled as ≡ should have a positive label in the original dataset. Yet, there are 0.9% ≡ pairs in MRPC and 13.8% ≡ pairs in QQP that have a negative label in the original dataset. Our further analysis shows that some are caused by original labelling issue and some are caused by classifier noise. We do find that QQP contains a fair bit of false negative samples. We suspect this could be caused by relying on Quora users to indicate if questions are a duplicate. Since users are not expected to have seen all questions, the dataset is bound to have relatively high false negative samples. We also find that both the MNLI and SNLI classifier tend to make a wrong prediction in sentences with an ambiguous pronoun reference. An interesting example is the sentence pair from MRPC with a negative label: "NBC will probably end the season as the second most popular network behind CBS, although it's first among the key 18-to-49-year-old demographic." and "NBC will probably end the season as the second most-popular network behind CBS, which is first among the key 18-to-49-year-old demographic." Both the original and swapped pairs are classified as "entailment".

We also expect a very small percentage of positive labels in the negation (∧) class. Yet, the results show that 39.6% of negation class pairs have positive labels in MRPC. Manual inspection reveals that most of those sentence pairs contain numeric information. Apparently, paraphrase corpus and language inference corpus have different interpretations on two sentences and differ only in numeric information. An example pair such as " They remain 40 percent below the levels prior to February 's initial overstatement news" and "The stock remains 43 percent below levels prior to the February overstatement news " is annotated as positive in MRPC but is classified as "contradiction" by the language inference classifiers.

5.2. String Property Analysis

Table 11 shows the average word length difference between the first sentence and second sentence of each class. The sentence pairs in paraphrase datasets have very close word length, as shown in Table 11a. In both datasets, the second sentence length is slight longer than the first sentences in the negative class. For the positive class, almost all the sentence pairs share the same length. As for the language inference datasets, most sentence 1 in the sentence pairs tend to be relatively longer than sentence 2. Compared to SNLI, MNLI has a much bigger length difference in the sentence pair samples. However, for both datasets, the length difference in each class is negligible.

Table 11. Sentence pair length difference.

(**a**) Paraphrase Datasets

Data	0	1	All
QQP	−0.4	0	−0.3
MRPC	−0.1	0	0

(**b**) Language Inference Datasets

Data	Contradiction	Neutral	Entailment	All
MNLI	12.3	11.1	12.5	12
SNLI	5.8	5	6.6	5.8

(**c**) Fine-grained Labelled Datasets

Data	Negation (∧)	Entailment (⊏)	Elaboration (⊐)	Equivalence (≡)	Independence (#)
QQP	0	3.3	−3.6	0	1
MRPC	0	3.1	−3	0	0.1
MNLI	10	16	2.4	4.3	13.1
SNLI	5.6	7.3	−1	0.8	6

On the contrary, the length difference of each fine-grained label is substantial. For the bidirectional entailment class (≡), all datasets except MNLI has a length difference close to 0. In terms of the elaboration class (⊐), the second sentences tend to have longer word length than the first sentences except for MNLI. In contrast to elaboration, the entailment class (⊏) has longer first sentences. Due to the considerable difference in sentences length of MNLI, the average length of sentence1 is longer in all fine-grained classes. However, the difference of each class can still be reflected in Table 11c.

The bidirectional entailment paraphrases preserve all the amount of information in the given sentences. Thus, the sentence pair should theoretically have the same sentence length. On the other hand, unidirectional entailment paraphrases, including elaboration class and the entailment class, exhibit a certain amount of content loss [11]. A typical example of content loss by deletion can be observed from the following sentence pair:

- Yesterday I went to the park.
- Yesterday I went to Victoria Park.

This explains the sentence length difference of the two unidirectional entailment class in Table 11c.

In addition to sentence word length, we also examine the Rouge-L score of each fine-grained class. Rouge-L takes into account sentence level structure similarity naturally and identifies the longest co-occurring in sequence n-grams automatically [22]. Table 12c shows the Rouge-L F1 score of each fine-grained class. For paraphrase datasets, the positive class has a higher Rouge-L score. MRPC sentences are mainly extracted from news while QQP contains duplicated question pairs in the knowledge sharing forum Quora. Thus, the sentences of MRPC are normally longer than QQP. This explains why MRPC tends to have higher Rouge-L score in all classes compared with QQP. As for the language inference datasets, the entailment class has a higher score than other classes.The MNLI contradiction class has a higher score compared with neutral class while the SNLI contradiction class has a lower score.

Table 12. Sentence pair Rouge-L F1 score.

(**a**) Paraphrase Datasets

Data	Negative	Positive
QQP	45.26	64.65
MRPC	59.22	71.28

(**b**) Language Inference Datasets

Data	Contradiction	Neutral	Entailment
MNLI	35.59	32.8	44.48
SNLI	33.47	38.35	45.16

(**c**) Fine-grained Labelled Datasets

Data	Negation (∧)	Entailment (⊏)	Elaboration (⊐)	Equivalence (≡)	Independence (#)
QQP	47.79	60.09	58.98	**69.17**	37.64
MRPC	61.08	71.72	71.69	**75.92**	61.78
MNLI	37.02	40.01	47.2	**53.97**	29.58
SNLI	34.57	43.79	58.83	**62.09**	33.18

In terms of fine-grained labels, the paraphrase classes, including entailment, elaboration and equivalence, have a higher score than the other fine-grained classes. Among the paraphrase classes, the equivalence class has the highest Rouge-L score of all the original labels and the other fine-grained classes. As Table 11c suggests, the elaboration class a has much lower sentence length difference than the entailment class in language inference data. As a result, the elaboration of language inference data performs better than the entailment class. On the contrary, both classes of paraphrase datasets have similar score.

6. Generation Experiment

In this section, we demonstrate that the newly labelled dataset can be used to train sentence generation models' more specific requirements. We are able to train a model that generates an equivalent paraphrase and models that generate paraphrases with a specific entailment direction.

6.1. Experiment Models

Our generation models are fine-tuned on the large version of pre-trained **T5** text-to-text transfer transformer model [23]. We train three types of generators: equivalence generator, forward entailment generator and reverse entailment generator. Each model is trained with the corresponding class examples in a relabelled dataset. In total, we trained nine generators using training data from relabelled QQP, SNLI and MNLI datasets. Table 13 shows the number of sample pairs for training, validation and testing, respectively.

Table 13. Training sample size.

Data	Class 1 (⊏)		Class 2 (⊐)		Class 3 (≡)	
	Train	**Test and Val.**	**Train**	**Test and Val.**	**Train**	**Test and Val.**
MNLI	86,950	2253	22,983	551	41,787	1117
SNLI	164,285	3183	21,679	387	18,251	311
QQP	47,710	5440	45,688	5055	76,806	8373

We did not create generation models for class 0 (negation) and class 4 (alternation) because both classes a have broad definition and would need a lot more training samples to generate a meaningful sentence with the correct relation. As shown in Figure 1, we add the prefix "paraphrase" to the first sentence as an input and the second sentence is used as the output.

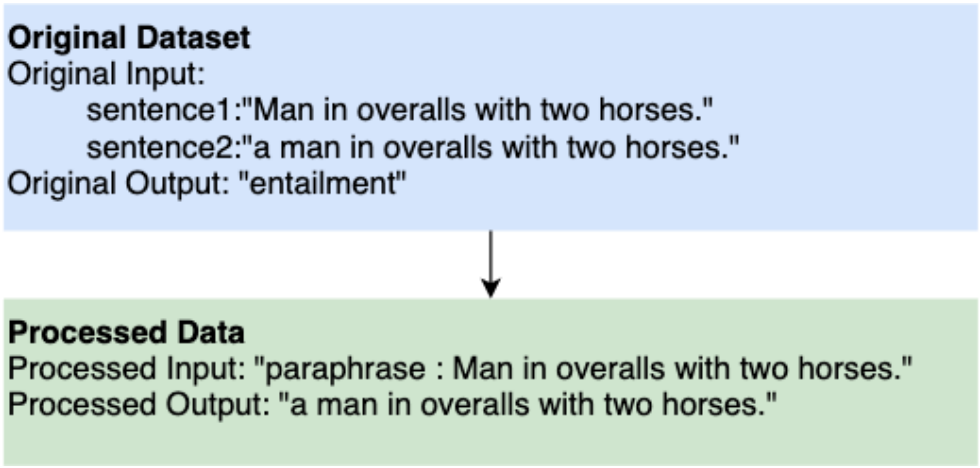

Figure 1. Processed input for text generation.

Hyperparameters: During the pre-processing stage, the maximum token length is set to 128 to get an equal length of input vectors. This means that the input sequence with less than 128 tokens will be padded and the input sequence with more than 128 tokens will be truncated.

Adam with weight decay is used as the optimiser as it is proven as the fastest way to train neural nets [24]. The learning rate is 2×10^{-5} and the weight decay rate is set to 0.01. The maximum number of an iteration is set to 20 with the minimum improving value 0.01 of the loss function. The mini batch is set to 4 to cope with the constrain of the graphics processing unit (GPU) memory size.

To get the optimal result, we use beam search for text generation. We adopt parameters from T5 [23]. Specifically, we use K = 50 for TopK and a beam width of 5 with 1.0 length penalty for beam search. The generated length is limited to 128.

The training process utilises GPU Tesla V100-SXM2-16GB provided by Google colab. The Huggingface implementation version of the pre-trained transformer model is used for the downstream tasks. All the text generation models are stopped early by reaching the minimum loss gap. Table 14 shows the training time and respective losses. All of the models have a similar training and dev loss, indicating that the models are neither under fitting nor over fitting. It is consistent in different corpora that among all of the classes, the equivalent paraphrases (≡) model tends to have the lowest training loss.

Table 14. Training results.

Data	Training Time	Epoch	Train Loss	Dev Loss
MNLI 1 (⊏)	116 min	2	0.140	0.139
MNLI 2 (⊐)	62 min	4	0.195	0.214
MNLI 3 (≡)	87 min	3	0.09	0.106
SNLI 1 (⊏)	248 min	2	0.08	0.08
SNLI 2 (⊐)	49 min	3	0.164	0.161
SNLI 3 (≡)	20 min	2	0.108	0.09
QQP 1 (⊏)	116 min	3	0.07	0.07
QQP 2 (⊐)	221 min	5	0.146	0.178
QQP 3 (≡)	148 min	3	0.05	0.06

6.2. Generator Results

We first examine the textual property of the generated text, including sentence length, dependency tree height and Rouge-1 score. Then, we classify the generated sentence pair to test whether the text generated has the intended similarity direction. Table 15 shows the result sentence length of the input sentence, target sentence and generated sentence. As mentioned in the previous chapter, the sentence length difference in input sentence and target sentence of the unidirectional entailment class is substantial, while the equivalence class has a very small length difference. Likewise, the generated sentences behave similarly to the target sentence since the difference in the target sentence and generate sentence is insignificant. Thus, we can say that word length wise, the models trained on the fine-grained data can generate text with specific direction.

Table 15. Sentence length comparison.

Data	Input Sentence	Target Sentence	Generated Sentence
MNLI 1 (⊏)	26.3	11.6	13.8
MNLI 2 (⊐)	10.3	12.3	12.6
MNLI 3 (≡)	11.8	10.7	10.9
SNLI 1 (⊏)	15.8	7.4	10.1
SNLI 2 (⊐)	10	11.6	12.3
SNLI 3 (≡)	11.3	10.4	11.3
QQP 1 (⊏)	13.1	9.8	9.9
QQP 2 (⊐)	10	13.4	12.3
QQP 3 (≡)	9.9	9.9	9.6

Table 16 shows the dependency tree height difference between the input and generated sentence of each fine-grained class. Theoretically, paraphrase pairs are likely to have a similar dependency parse tree. This reflects in the equivalence class, where the input and output sentences have a very close dependency height (the values for equivalence class are highlighted in the table). On the other hand, the unidirectional entailment class shows a certain level of content loss. This also reflects in the table, where in the entailment class input sentence has a deeper dependency level.

Table 17 shows the Rouge-1 score of the generated texts of the nine generators. The other Rouge results show a similar trend. Overall, equivalence generators (class 3) have the highest Rouge score across all of the datasets and between the two decoding methods. This is as expected since equivalence is the most strict relation among the three and there might be smaller variations on the generated text. We also found that generators trained on QQP and SNLI have higher a Rouge score than generators trained on MNLI. Thwe MNLI

dataset contains a large set of language genres, which could lead to more variations in the generated text, making it deviate more from the target text.

Table 16. Average dependency tree height difference between input and output sentences.

Data	Entailment (⊏)	Elaboration (⊐)	Equavalence (≡)
MNLI	1.41	−0.79	**0.2**
SNLI	1.61	−0.3	**0.09**
QQP	0.71	−0.59	**0.12**

Table 17. Generated text Rouge-1.

Data Sets	Refined Labels	Precision	Recall	F1 Score
QQP	1 (⊏)	61.80%	79.95%	68.76%
	2 (⊐)	72.69%	58.03%	63.82%
	3 (≡)	**79.5%**	**81.23%**	**79.96%**
SNLI	1 (⊏)	53.92%	80.78%	62.52%
	2 (⊐)	**92.74%**	73.85%	81.57%
	3 (≡)	90.15%	**90.24%**	**89.87%**
MNLI	1 (⊏)	49.26%	**82.07%**	59.05%
	2 (⊐)	69.76%	53.96%	53.96%
	3 (≡)	**71.03%**	77.80%	**73.40%**

Rouge scores only measure the syntax properties of the generated text. We next build several RoBERTa classifiers to check the relation between the source and the generated sentences. Table 18 shows results on three corpora. It is clear that all generators generate sentences in the desirable class in most cases. In all of the datasets, at least 80% of the generated sentences are classified as the desirable class. The percentage of correct relation is above 90% in many generators. Table 19 shows the sample source and generated sentence pairs from different datasets. We underline both the syntactic and semantic difference in sentence pairs. Note that all generator results used beam search decoding with a beam width of 4 and 0.6 length penalty. This is based on parameters reported in the original T5 paper. We also experimented top-k decoding with k = 50. The top-k decoding generates inferior results in general.

Table 18. Fine-grained paraphrase label percentage of generated text.

Datasets	Generator Class	Class 0 Percent	Class 1 Percent	Class 2 Percent	Class 3 Percent	Class 4 Percent
QQP	1 (⊏)	0%	**84.8%**	0%	14.7%	0.4%
	2 (⊐)	0.6%	0.4%	**82.6%**	16.0%	0.4%
	3 (≡)	0%	1.8%	0.4%	**97.7%**	0.2%
SNLI	1 (⊏)	0.1%	**97.5%**	0%	2.3%	0%
	2 (⊐)	1.1%	0.2%	**90.3%**	8.2%	0.1%
	3 (≡)	0%	4.4%	0.3%	**95.2%**	0%
MNLI	1 (⊏)	0.4%	**88.8%**	0%	10.5%	0.3%
	2 (⊐)	2.7%	0.7%	**80.2%**	15.9%	0.4%
	3 (≡)	0%	4.0%	0.7%	**94.9%**	0.3%

Table 19. Generated text samples.

Datasets	Sentence Relations	Source Text	Generated Text
QQP	Forward Entailment	What are the three places that anyone should visit at least once in their life, and why?	What are the places that anyone should visit at least once in their life?
	Reverse Entailment	What is a conspiracy theory that turned out to be real?	What is the most plausible conspiracy theory that turned out to be real?
	Equivalence	What are some ways to become an entrepreneur?	How can I become an entrepreneur?
MNLI	Forward Entailment	In front of the church tower is a Statue of St. Francis Xavier, its right arm broken off during a storm.	St. Francis Xavier's arm was broken off during a storm.
	Reverse Entailment	The woman was Nema.	Nema was the woman who was a nurse.
	Equivalence	What they owe us is an admission that their professed faith in term limits was phony in the first place.	They owe us an admission that their professed faith in term limits was phony in the first place.
SNLI	Forward Entailment	A man wearing a white helmet is rock climbing.	A man is rock climbing.
	Reverse Entailment	Four dogs in a grassy area.	Four dogs are playing in a grassy area.
	Equivalence	Three men in a foreign country sorting limes.	Three men sort limes in a foreign country.

7. Conclusions

Many natural language processing tasks involve paired sentence data. A single sentence is usually packed with lots of information, making it difficult to come up with a set of standard relations backed by rigorous logic and mathematics properties. We discovered that the paraphrase and entailment sentence relations defined in different benchmark tasks overlap with each other. This gives us the unique opportunity to extract fine-grained and cleaner relations from the existing datasets. Our new approach relies on the general properties of symmetric and asymmetric relations as well as the fact that a single class of current dataset contains sentence pairs belonging to multiple fine-grained relations. Our proposed relabelling approach produced a number of datasets with fine-grained paraphrase labels. They would enrich the existing benchmark corpora and would help in building general sentence encoding models as well as text generation models. The approach we take and the general rules defined can be applied in another dataset to generate fine-grained labels. Moreover, the relabelling process also reveals many useful features and properties of the current paraphrase and language inference dataset. Such properties help to provide insights on model performance as well as design optimisation.

Author Contributions: Conceptualisation, Y.Z. and X.H.; methodology, Y.Z.; software, X.H.; resources, X.H.; writing—review and editing, Y.Z. and V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study use several publicly available data sets with with links given in their respective reference. The data sets with new labels are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dolan, W.B.; Brockett, C. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005), Jeju Island, Korea, 14 October 2005.
2. De Beaugrande, R.A.; Dressler, W.U. *Introduction to Text Linguistics*; Longman: London, UK, 1981; Volume 1.
3. Shinyama, Y.; Sekine, S.; Sudo, K.; Grishman, R. Automatic paraphrase acquisition from news articles. In Proceedings of the HLT, San Diego, CA, USA, 2002; Volume 2, p. 1.
4. Barzilay, R.; Lee, L. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv* **2003**, arXiv:cs/0304006.
5. Bhagat, R.; Hovy, E. What is a paraphrase? *Comput. Linguist.* **2013**, *39*, 463–472. [CrossRef]
6. Qiu, L.; Kan, M.Y.; Chua, T.S. Paraphrase recognition via dissimilarity significance classification. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 18–26.
7. Radev, D.R.; Hovy, E.; McKeown, K. Introduction to the special issue on summarization. *Comput. Linguist.* **2002**, *28*, 399–408. [CrossRef]
8. Elhadad, N.; Sutaria, K. Mining a lexicon of technical terms and lay equivalents. In Proceedings of the Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, 29 June 2007; pp. 49–56.
9. Tomuro, N. Interrogative reformulation patterns and acquisition of question paraphrases. In Proceedings of the Second International Workshop on Paraphrasing, Sapporo, Japan, 11 July 2003; pp. 33–40.
10. Rus, V.; McCarthy, P.M.; Graesser, A.C.; McNamara, D.S. Identification of sentence-to-sentence relations using a textual entailer. *Res. Lang. Comput.* **2009**, *7*, 209–229. [CrossRef]
11. Vila, M.; Martí, M.A.; Rodríguez, H. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open J. Mod. Linguist.* **2014**, *4*, 205. [CrossRef]
12. Iyer, S.; Dandekar, N.; Csernai, K. First Quora Dataset Release: Question Pairs. 2017. Available online: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs (accessed on 20 November 2021).
13. Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv* **2017**, arXiv:1708.00055.
14. Pavlick, E.; Rastogi, P.; Ganitkevitch, J.; Van Durme, B.; Callison-Burch, C. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 425–430.
15. Bhagat, R.; Pantel, P.; Hovy, E. LEDIR: An unsupervised algorithm for learning directionality of inference rules. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 161–170.
16. Pavlick, E.; Bos, J.; Nissim, M.; Beller, C.; Van Durme, B.; Callison-Burch, C. Adding semantics to data-driven paraphrasing. 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), Beijing, China, 26–31 July 2015; pp. 1512–1522.
17. MacCartney, B. Natural Language Inference. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 2009.
18. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. *arXiv* **2015**, arXiv:1508.05326.
19. Williams, A.; Nangia, N.; Bowman, S.R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv* **2017**, arXiv:1704.05426.
20. Sukthanker, R.; Poria, S.; Cambria, E.; Thirunavukarasu, R. Anaphora and coreference resolution: A review. *Inf. Fusion* **2020**, *59*, 139–162. [CrossRef]
21. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
22. Lin, C.Y.; Och, F.J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2004; pp. 605–612.
23. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683.
24. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101.