Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09715-7

ORIGINAL PAPER

Semantic search as extractive paraphrase span detection

Jenna Kanerva[1] · Hanna Kitti[1] · Li-Hsin Chang[1] · Teemu Vahtola[2] · Mathias Creutz[2] · Filip Ginter[1]

Accepted: 13 December 2023

Abstract

In this paper, we approach the problem of semantic search by introducing a task of paraphrase span detection, i.e. given a segment of text as a query phrase, the task is to identify its paraphrase in a given document, the same modelling setup as typically used in extractive question answering. While current work in paraphrasing has almost uniquely focused on sentence-level approaches, the novel span detection approach gives a possibility to retrieve a segment of arbitrary length. On the Turku Paraphrase Corpus of 100,000 manually extracted Finnish paraphrase pairs including their original document context, we find that by achieving an exact match of 88.73 our paraphrase span detection approach outperforms widely adopted sentence-level retrieval baselines (lexical similarity as well as BERT and SBERT sentence embeddings) by more than 20pp in terms of exact match, and 11pp in terms of token-level F-score. This demonstrates a strong advantage of modelling the paraphrase retrieval in terms of span extraction rather than commonly used sentence similarity, the sentence-level approaches being clearly suboptimal for applications where the retrieval targets are not guaranteed to be full sentences. Even when limiting the evaluation to sentence-level retrieval targets only, the span detection model still outperforms the sentence-level baselines by more than 4 pp in terms of exact match, and almost 6pp F-score. Additionally, we introduce a method for creating artificial paraphrase data through back-translation, suitable for languages where manually annotated paraphrase resources for training the span detection model are not available.

Keywords Paraphrase retrieval · Finnish · Semantic search · Paraphrasing

✉ Jenna Kanerva
jmnybl@utu.fi

[1] TurkuNLP, Department of Computing, University of Turku, Turku, Finland

[2] Department of Digital Humanities, Faculty of Arts, University of Helsinki, Helsinki, Finland

Published online: 01 February 2024

1 Introduction

With the existence of large, pre-trained language models, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), or T5 (Raffel et al., 2020), numerous NLP tasks requiring deep language understanding have recently gained promising results. For example, in natural language inference and question answering such models have helped to substantially narrow down the gap between human and model performance (see e.g. Sun et al. (2021) or Raffel et al. (2020)). One task clearly requiring deep language understanding is semantic search, where the objective is to retrieve the relevant knowledge based on the actual meaning of the search query rather than its surface form only. For example, when querying using the phrase *the dimensions of Volkswagen Transporter* also documents mentioning the paraphrased versions *VW Transporter: size* or *the length, width and height of VW Transporter* should be considered relevant, and a good search engine would be expected to prioritize these over the bare keyphrase of *Volkswagen Transporter*.

A full system for semantic search typically consists of retriever and reader components, where the retriever receives a search query q and a large candidate document collection, and returns the document d containing the relevant information, while the reader receives the search query q and the retrieved document d, and extracts the relevant span from d either containing the paraphrased version of q or the correct answer for it depending on the query type. Therefore, the reader can be seen as in part carrying out *paraphrase detection*, i.e. identifying statements equivalent in meaning with the search query but differing on the surface level. While retriever and reader models are actively studied topics in the area of question answering and machine reading comprehension (see e.g. Chen et al. (2017); Zeng et al. (2020); Qu et al. (2021)), current work in paraphrase detection and retrieval has almost uniquely focused on sentence-pair classification or sentence-level retrieval approaches, rather than span extraction. In this work, we take a novel span-level extraction approach to the paraphrase detection task: Given a document and a segment of text as a query, the task of the model is to identify a paraphrase of the query from the document. Therefore, our work can be seen as concentrating on the overlap of the paraphrase detection task and search systems reader component methodology.

Recently, a large-scale corpus of Finnish paraphrases, the Turku Paraphrase Corpus (Kanerva et al., 2023), became available. The paraphrase pairs in the corpus are manually extracted from pairs of related documents, forming annotated examples where the document context of both members of the paraphrase pair is known (for illustration, see the left side of Fig. 1). This very property is to the best of our knowledge unique to this corpus and in turn allows us to take this novel paraphrase span detection approach to semantic search, where one of the paraphrases act as a search query, while the objective is to identify the span from the given context document constituting the paraphrased pair for it. The primary advantage of using span detection, as opposed to the conventional approach of classifying sentence pairs or computing their pairwise similarity, is the ability to

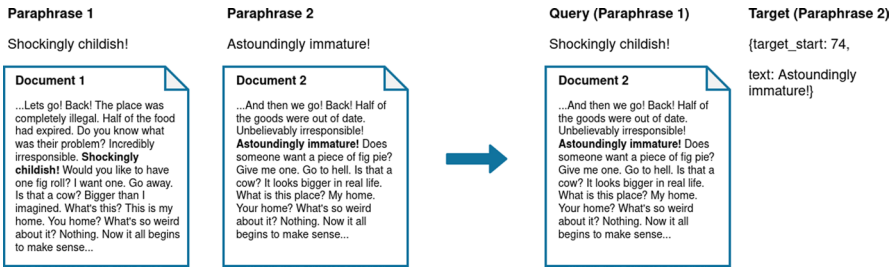

Fig. 1 On the left side is an illustration of one paraphrase pair from the Turku Paraphrase Corpus, and on the right side is the same paraphrase pair turned into the span detection framework as used in this work

easily extract any part of the target document, not just predefined units such as lines or sentences. Additionally, the models are informed about document context and, at least in theory, able to extract also pairs which, if taken out of context, would not be judged as paraphrasing each other.

In summary, the key contribution of this work is in introducing the novel paraphrase-span detection task, where the span detection models are shown to outperform several sentence-level retrieval baselines on the Finnish paraphrase data. Specifically, we show that widely adopted sentence-level paraphrase approaches are not sufficient for applications where the retrieval targets are not guaranteed to be sentences. Additionally, we introduce a straightforward method of generating artificial paraphrase data through back-translation, allowing training span models also for languages where manually annotated paraphrases-in-context data is not available. Finally, we carry out an extensive error analysis to understand the prediction capabilities of the span detection model.

2 Related work

2.1 Paraphrase

In NLP, different paraphrase related tasks include detecting, retrieving or generating paraphrased versions of a given text span. Numerous paraphrase corpora, e.g. Quora Question Pairs[1], Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005), and PARADE (He et al., 2020), have been released for these purposes, each including labeled sentence-like text pairs supporting mainly paraphrase classification. Paraphrase retrieval is typically approached using large monolingual corpora and performing one-to-many (find a paraphrase for the given text span) or many-to-many (find all paraphrase pairs from the text collection) sentence similarity comparison between calculated sentence-level embeddings (see e.g. Vrbanec and Meštrović (2020)), potentially including document level heuristics in order to restrict the search space into comparable documents. To our knowledge, besides the Turku Paraphrase

[1] data.quora.com/First-Quora-Dataset-\Release-Question-Pairs.

Corpus, there is no other paraphrase corpora available where the original document context would be available for the paraphrase pairs, and thus directly supporting the paraphrase span detection task.

2.2 Question answering

In extractive question answering, the system is given a question posed in natural language together with a background document, and the task is to extract the span of the correct answer from the document. Span detection is a common approach for extractive question answering, naturally supporting extracting an answer segment of any length. There are both mono- and multilingual question answering datasets available. SQuAD (Rajpurkar et al., 2016) is an English QA dataset including approx. 100,000 examples where the context document has an answer for the given question. In its second release (SQuAD v2), also unanswerable questions are included (Rajpurkar et al., 2018). Some multilingual QA corpora include e.g. XQuAD (Artetxe et al., 2020), TyDiQA (Clark et al., 2020), and MKQA (Longpre et al., 2020), the latter two including also Finnish examples. Even though the task setup used in this work resembles the QA task, the objective is different. While in QA the system is expected to return an answer for the question, in paraphrase retrieval it returns a semantically equivalent segment from the background document.

2.3 Semantic textual similarity

In the semantic textual similarity (STS) task, each sentence pair is annotated with a similarity score typically ranging from 0 to 5, where lower scores mean unrelated or related sentences, while higher scores are for partially or fully equivalent sentences, with the highest score typically indicating the sentences being completely equivalent in meaning. The annotations in STS and paraphrasing tasks are highly related (Gold et al., 2019) but not necessarily completely interchangeable between different datasets as the definition of a paraphrase or relatedness may not be fully equivalent. Similar to paraphrase datasets, most of the STS datasets include pairs of approximately sentence-long text snippets together with the annotated degree of similarity, therefore supporting the setting of a sentence-pair classification task without contextual information (see e.g. Agirre et al. (2016); Cer et al. (2017)). However, a recent dataset of Sido et al. (2021) includes similarity annotations of Czech sentence pairs in document context, thus to our knowledge being the first STS dataset which could directly be applied to span detection modelling.

3 Data

The Turku Paraphrase Corpus[2] consists of paraphrase pairs manually extracted from pairs of related documents with high probability for naturally occurring paraphrases. The paraphrases can be anything between short phrases to several

[2] Newest data release available at: https://github.com/TurkuNLP/Turku-paraphrase-corpus

Table 1 Dataset sizes after converting the original paraphrase data into the span detection framework, where Setup 1 includes only retrievable examples, while Setup 2 includes both retrievable and irretrievable examples

Section	Setup 1 Examples	Setup 2 Examples
Train	138,706	140,848
Devel	17,702	17,930
Test	17,564	17,810
Total	**173,972**	**176,588**

sentences long segments, and the position in the respective source documents is preserved to obtain contextual information. Most of the pairs are obtained from independent subtitle versions of the same movie or TV episode. Subtitles thus constitute the primary domain of the data, while a small portion is extracted from other domains, including news articles, discussion forum messages as well as university exercises and essays. Furthermore, each paraphrase pair is manually categorized in a scheme distinguishing paraphrases primarily by the degree of their context independence. In Fig. 1 we illustrate one paraphrase pair from the original corpus, as well as its transformation into the span detection setting used in this work.

The corpus has two categories of examples of interest for this study: 86,986 positive examples of naturally occurring paraphrases in their respective document contexts and 1,308 negative examples of pairs in their document contexts that are semantically similar but not mutual paraphrases. These constitute 84% of the corpus. The remaining 16% are unsuitable for this study as they either do not have document contexts for various reasons, or are manually edited and therefore no longer naturally fitting their contexts.

As the paraphrase pairs are not directional in the same manner as for example question-answer examples are, and the two paraphrases are always extracted from two distinct context documents, each pair produces two distinct examples in the span detection task, resulting in a total of 173,972 distinct positive and 2,616 distinct negative examples. The data statistics are summarized in Table 1 in terms of train, development and test sets, following the dataset split provided in the original corpus.

We pursue two different task setups: (1) Retrievable paraphrases formed from the positive examples, where for all examples a valid paraphrase is guaranteed to exist in the context. The setup is similar to SQuAD v1 in question answering. (2) Including the 2,616 negative examples as irretrievable paraphrases, requiring the model not only to find a valid paraphrase, but also being able to determine when there is not a valid paraphrase present in the context. The setup is similar to SQuAD v2 in question answering.

4 Experiments

4.1 Paraphrase-SD model

As the span detection model (referred to as *Paraphrase-SD* throughout the paper), we use the HuggingFace library's (Wolf et al., 2020) implementation of the standard modelling approach to the question answering task. The model is illustrated in Fig. 2 and described below. Given a query phrase and a document, separated by the [SEP] token, the model detects the span in the document as follows: Each subword encoded by the BERT model is classified by two classification layers, one for predicting the start position of the span and one for predicting the end position. The output of the model is then the span which maximizes the sum of the logits for its opening and closing subword. In Setup 2 that includes also irretrievable examples, the model is trained to predict the [CLS] token as both start and end position of the span for negative examples, thus in practise returning an empty span (null prediction). We use the Finnish *FinBERT* (Virtanen et al., 2019) language model as the encoder.

Many of the documents are longer than the maximum sequence length of 512 subwords in the BERT model. We slice the documents (with overlap of 128 subwords) into segments, which form independent examples. Each of these examples thus consists of the query phrase which is never sliced nor truncated, and a slice of the document. Predictions for these examples are subsequently merged into a single prediction for the whole document, as follows: In Setup 1 including only retrievable paraphrases, the span with the highest aggregated score out of all possible spans over all slices of the document is chosen. However, this necessary slicing interferes with Setup 2. When there are multiple document slices, the model is likely to give a highly confident null prediction for all slices not including the target span. When aggregating the scores across all slices of the document, these confident null predictions would dominate the output. Noting that all document slices give some probability for the null prediction, the final null prediction score can be obtained by taking the minimum value (least confident null prediction) across all document slices,

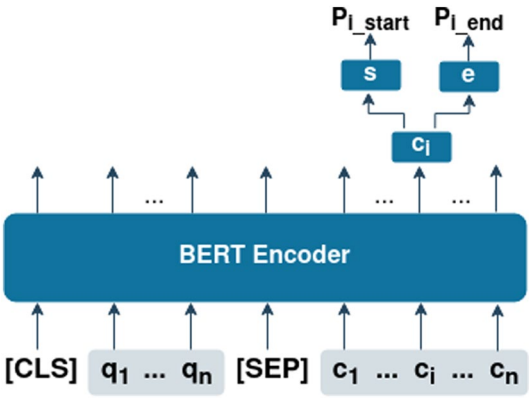

Fig. 2 Illustration of the span detection model when predicting start and end probabilities for the context token c_i

approximating the null prediction value obtained for the full document at once. That score is then compared against the most confident span predictions selecting the span with the highest overall value as the final prediction, therefore predicting an empty span if even the least confident null prediction has higher value than the most confident non-empty span.

The weights of the pre-trained FinBERT language model are fine-tuned together with the two task specific classification layers during training. We performed a grid search separately for Setup 1 and Setup 2 in order to find optimal hyperparameters. Trialed hyperparameters were batch sizes 8, 16 and 32, learning rates 5e-5, 3e-5 and 2e-5 and epochs 2 and 3 on development section of the data. For all experiments with Setup 1 we use batch size 32, learning rate 3e-5 and the model is trained for two epochs. Respectively, for all the experiments with Setup 2 the hyperparameters are: batch size 16, learning rate 2e-5 and two epochs. The source code for all our experiments is available at https://github.com/TurkuNLP/paraphrase-span-detection.

4.2 Baselines

We compare the Paraphrase-SD model with several baselines typically applied to paraphrase retrieval. Our baselines are based on embedding similarity, where for each paraphrase in the evaluation data, the most similar sentence in the target document is retrieved based on the cosine similarity of embeddings obtained using three different baseline models, tf-idf, BERT and fine-tuned Sentence-BERT. In addition to the embedding based methods, we also evaluate using BM25, a ranking function typically used in information retrieval and search engines. Similarly to other baselines, the BM25 is also applied on sentence-level, approximating the best predefinable target unit for retrieval.

Both BM25 and tf-idf are straightforward methods measuring lexical similarity. For BM25, for which slightly varying implementations are introduced, we use the Okapi BM25 version as defined e.g. in Wikipedia.[3] The two parameters (k and b) are optimized on the development data. For tf-idf, we use tf-idf weighted vectors with the union of character n-grams of lengths 2, 3 and 4 created inside word boundaries. The n-gram vocabulary was induced on the training data only. In the BERT baseline, the embedding for each sentence is calculated as the average of token embeddings obtained from the last hidden layer of the FinBERT model without any fine-tuning. While BM25, tf-idf and BERT are unsupervised methods, we yet evaluate embedding models directly fine-tuned into semantic relateness using the Sentence-BERT (SBERT) approach (Reimers & Gurevych, 2019). We test two models, multilingual SBERT[4] (mSBERT) trained on English paraphrases as well as parallel translation pairs from over 50 languages, and Finnish SBERT[5] (fiSBERT) trained on the Turku Paraphrase Corpus.

[3] https://en.wikipedia.org/wiki/Okapi_BM25

[4] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

[5] https://huggingface.co/TurkuNLP/sbert-cased-finnish-paraphrase

One notable advantage of the Paraphrase-SD model is its ability to return any text segment from the background document. All of our baselines, on the other hand, are limited to sentence level predictions, in order to avoid embedding all possible document segments of any length, which would be highly impractical. However, as the paraphrases in the Turku Paraphrase Corpus are not strictly limited to sentence boundaries, with about 25% being longer or shorter than a sentence, the baseline approaches incur an inevitable loss. To assess its magnitude, we will report also the oracle performance, corresponding to returning the one sentence from the document that is most overlapping with the true target span. We are not aware of any baseline methods capable of returning a span of arbitrary size in reasonable compute time, however in Sect. 5.2 we provide also a comparison of the main and baseline methods when limiting the dataset to sentence-level targets only, therefore eliminating this disadvantage of the baseline approaches. In addition to these, in Sect. 5.3 we yet run an exhaustive baseline experiment, where we systematically generate all possible token-level spans of any length in order to evaluate whether the baseline methods would be able to return accurate non-sentential targets if not taking into account the computational limitations of such approach.

4.3 Paraphrase-SD through back-translation

Up to this point, we relied on the fact that the Turku Paraphrase Corpus enables our approach by containing paraphrases in their context. Such a dataset is, to the best of our knowledge, currently available only for Finnish. In this section, we explore a straightforward heuristic approach based on a form of back-translation (Sennrich et al., 2016), allowing the application of the Paraphrase-SD model also in absence of such a manually annotated corpus.

We take an approximate of 60K Finnish subtitle files from the same subtitle domain as in the Turku Paraphrase Corpus but which were unused in the original data. We split the files into shorter text segments yielding 260K documents. Additionally, we have acquired approximately 200K Finnish documents from the Reddit discussion forum to accompany the data. As illustrated in Fig. 3, for each of these documents, we randomly sample one sentence from any position in the document to act as a target sentence whose span is to be retrieved from the document in the paraphrase span detection task. We remove examples consisting of target sentences longer than 100 tokens to reduce unnecessarily noisy examples. Finally, we use back-translation to generate an assumed paraphrase for each sampled target sentence that the Paraphrase-SD model is supposed to retrieve. We translate the original Finnish target sentences into English, and back into Finnish using pre-trained translation models from the OPUS-MT project (Tiedemann & Thottingal, 2020). We decode the translations using beam search with a beam size of 6 and a length normalization term of 0.6 in both directions. Based on translation probabilities, we select the most probable back-translated sentence for each source sentence to act as a paraphrase of the original. The back-translated sentence is always used as the query phrase, while the original sentence in its context acts as the retrievable target span. While we currently implement the back-translation using sentence-level targets only, the

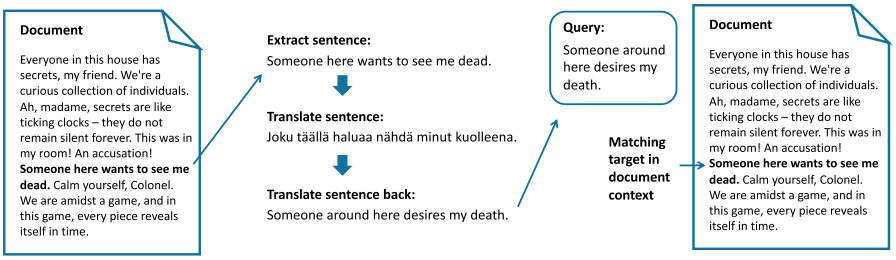

Fig. 3 Back-translation technique, which creates queries for targets in context. For illustration purposes, the text is here in English and the extracted sentence is translated to Finnish and back to English. In reality, the text is in Finnish, and translation is carried out via English back to Finnish

same could be tested with targets of any length by occasionally sampling also two consecutive sentences or segments shorter than a sentence (e.g. individual clauses or phrases) for back-translation. However, we leave this as a future work.

The back-translated data is used to train the span detection model using the same hyperparameters as with the original model. The back-translated data was randomly sampled to the same size as the original training data, however before sampling we removed examples where the back-translation produced an identical sentence compared to the target span, an empty sentence, or a sentence longer than 380 subwords.[6] Since the back-translated data contains only retrievable paraphrases, it is used in the Setup 1 experiments only. The final size of the training data for the back-translation model is 138,706 examples.

5 Results

The main results are summarized in Table 2 using two evaluation metrics: *exact match (EM)*, which measures the percentage of predictions that match the gold segment exactly disregarding only punctuation characters and casing, and *F-score*, which measures the average token-level overlap between the prediction and the gold segment, when segments are treated as bag-of-tokens disregarding punctuation characters and casing. These are the same metrics as typically used in the evaluation of extractive QA systems. All results are reported using the test section of the Turku Paraphrase Corpus (over 17,000 examples for both setups, see Table 1).

Our main model, Paraphrase-SD trained with the Turku Paraphrase Corpus, outperforms all baselines in both setups with a clear margin, receiving EM 88.73 and F-score 94.31 for Setup 1 (only retrievable paraphrases), and EM 84.37 and 89.52 F-score for Setup 2 (retrievable and irretrievable paraphrases). The second best performing model, Paraphrase-SD model trained with back-translation data, sees about − 17.4 pp decrease in EM compared to the main model in Setup 1. All sentence-level

[6] Sentences longer than 380 subwords were filtered out due to preserving enough space for the document in the model's input.

Table 2 The main results for the Setup 1 including only retrievable paraphrases as well as for Setup 2 where both retrievable and irretrievable paraphrases are used

Model	Setup 1		Setup 2	
	EM	F-score	EM	F-score
Sentence-level baselines				
BM25	49.53	64.64	48.85	63.76
TF-IDF	56.84	72.02	56.06	71.03
BERT	66.32	81.44	65.40	80.31
mSBERT	68.00	82.88	67.05	81.73
fiSBERT	64.72	79.03	63.83	77.94
Oracle	76.74	93.85	75.70	92.57
Paraphrase-SD				
Back-translation	71.32	85.07	–	–
Main model	**88.73**	**94.31**	**84.37**	**89.52**
Data augmentation				
+ Artificial irretrievables	–	–	82.35	87.11
+ Back-translation (random)	88.67	**94.48**	–	–
+ Back-translation (tf-idf most dissimilar)	88.38	94.24	–	–
+ Back-translation (tf-idf 0.35−0.66)	88.61	94.28	–	–

Results are reported in terms of exact match (EM) and token-level F-score on test section

baselines fall behind the back-translation, the best fine-tuned model (mSBERT) having − 20.7 pp decrease and the best unsupervised model (BERT) − 22.4 pp decrease in EM compared to the main model. The results are similar in terms of F-score, the back-translation and the two mentioned baselines (mSBERT and BERT) being − 9.2 pp, − 11.4 pp, and − 12.9 pp behind the main model. Both lexical baselines (BM25 and tf-idf) are clearly underperforming compared to other models. The sentence-level retrieval significantly harms the theoretical upper bound (oracle) of the baselines in terms of EM, and to a much lesser degree in terms of F-score, clearly demonstrating the intrinsic disadvantage of limiting retrieval to such pre-defined units in applications where the targets are not guaranteed to be sentences. Nevertheless, all baselines fall notably behind the oracle performance in both metrics, showing that the sentence-level retrieval is not the main limiting factor in the baseline performance. This will be further discussed in Sects. 5.2 and 5.3. By comparing the Setup 1 and Setup 2 results for the main model, we can see that including the irretrievable cases in the training data and asking the model to recognize when the correct paraphrase does not exist in the document decreases the performance of the Paraphrase-SD model. Naturally, the behavior is expected due to introducing a more difficult task setup.

Next, we will proceed with preliminary data augmentation experiments to explore if the performance of our primary model can be improved by introducing automatically generated additional training data, an approach frequently demonstrated to enhance the performance of supervised models in situations where data availability is the limiting factor. Subsequently, we will address the mismatch between our

span-level model and sentence-level baselines from two perspectives: first, by evaluating using only sentence-level targets, and second, by extending the sentence-level baselines to return spans of any length by iterating through all possible spans.

5.1 Data augmentation experiments

The Turku Paraphrase Corpus contains only a small fraction of non-paraphrase pairs, corresponding to irretrievable examples in our Setup 2. Therefore, we first experiment with increasing the small proportion of irretrievable examples in the training data by automatically creating artificial irretrievable training examples for the Setup 2 model. These are created from retrievable examples by simply removing the target span from the document. Each retrievable training example is thus introduced twice in the training data, once as a retrievable example and once as an artificial irretrievable example, resulting in total of 279,554 training instances with approximately 50/50 label distribution. Nevertheless, we find that the artificial irretrievables did not improve the model performance, being approximately -2pp worse than the original model on both metrics, mostly resulting from a noticed increase in false null prediction rate. This is not a particularly surprising finding, given that the evaluation data is not changed, causing a distribution mismatch between training and evaluation data. Since the removed target span no doubt leaves an unnatural artefact in the document, in the place where the target used to be, which the model can learn to recognize, the results on such artificially modified evaluation data would not have been reliable. The fact that the decrease in performance is quite limited, even though the training data distribution is substantially altered shows a surprising resilience of the model.

In the main results, we showed the span-detection model trained purely on back-translation data to exceed the performance of the SBERT, BERT and lexical baselines (row *Back-translation* in Table 2), indicating the back-translation as a viable option for training a span-detection-based retrieval model if manually curated training data for such model does not exist. A natural follow up question is whether the performance of our main model could be improved by enhancing our primary training data with artificially created back-translation examples, and therefore increasing the size of the training data available for the task. To investigate the hypothesis, we carry out preliminary data augmentation experiments, where we train three additional models with mixtures of original and back-translation training data using different sampling strategies. Each experiment includes exactly 138,706 back-translation examples matching the size of the original training set, and thus doubles the training data compared to the original Paraphrase-SD model. The first model uses a random sample of the back-translation data, the second model strives to include "interesting" examples with low lexical overlap obtained by sampling the most dissimilar query–target pairs in terms of tf-idf similarity, and the third model balances between too similar (trivial examples) and too dissimilar (likely including translation errors) by sampling mid-range examples using tf-idf similarities between 0.35–0.66. The tf-idf similarities are calculated using the same parameters as for the tf-idf sentence retrieval baseline.

The results (Table 2) show that our initial approach to combine the original training data from the manually annotated paraphrase corpus and the back-translation data did not improve the overall results. The best result was obtained with the randomly sampled back-translation data, exceeding the performance of the main model by a mere + 0.2pp. Our experiments of selecting more interesting back-translation examples did not yield positive results over the random selection. However, we see several limitations in our preliminary experiments. While we tested different strategies to select diverse examples from the back-translation data, we did not experiment with methods to create more diverse examples during the translation process, such as employing diverse beam search during translation (Vijayakumar et al., 2018)). Additionally, our back-translation data was generated by sampling sentence-level targets only and therefore not utilizing the span approach in its full power. Better results could be obtained by occasionally sampling also two consecutive sentences, or segments shorter than a sentence. However, to maintain the focus and scope of this paper, we do not proceed examining the multitude of possible strategies of sampling and incorporating the back-translation data. Nevertheless, given the positive outcome when compared to the baselines, more detailed experiments are clearly justified as a future work.

5.2 Sentence-level targets

As mentioned earlier, the Turku Paraphrase Corpus includes a significant amount of examples where the target is not a sentence (approx. 25% of the data), which in combination of the new task setup creates a situation where the sentence-level retrieval is not sufficient, causing a lower theoretical upper bound (oracle) for the sentence-level baseline methods not suffered by our Paraphrase-SD model. Note that we see this as a limitation of the existing baseline methods rather than the evaluation setup. To establish a comparison without the effect of non-sentential segments, we additionally evaluate on the subset of the data where both methods have a theoretical upper bound of 100%. To this end, we limit the test data to sentence-level by discarding all examples where the retrieval target is not a single, complete sentence, and compare which of the two approaches, sentence embeddings or span detection, gives better performance accuracy-wise if the objective is to retrieve one sentence from the target document paraphrasing the query. However, one must note that the span-level models were not retrained nor forced to return sentences, this time giving a slight advantage to the sentence-level baselines.

The results are shown in Table 3. Even in this evaluation setup, our span-level main model is visibly better than the baselines, outperforming the best baseline (mSBERT) by 4.5pp in terms of EM, and 5.9pp in terms of F-score. However, the difference of the span detection model trained purely on back-translation data as compared to the sentence-level baselines is now smaller, the mSBERT slightly outperforming the back-translation model in terms of EM, while the back-translation model is slightly better in terms of F-score. This indicates the back-translation approach being beneficial over the sentence-level retrieval on settings where the retrieval targets are not guaranteed to match sentence boundaries.

Table 3 Comparison of the paraphrase-SD model and the sentence-level baseline methods when limiting the test data to sentence-level targets only, and therefore simplifying the data in order to eliminate the disadvantage of the baseline methods

Model	Setup 1	
	EM	F-score
Sentence-level baselines		
BM25	64.19	67.25
TF-IDF	73.95	76.14
BERT	86.39	87.55
mSBERT	88.25	89.25
fiSBERT	84.06	85.19
Oracle	100.00	100.00
Paraphrase-SD		
Back-translation	87.87	90.60
Main model	**92.78**	**95.17**

5.3 Exhaustive baseline

To address questions whether lexical and embedding-based retrieval baselines would be able to return non-sentential segments if allowing them to do so, we yet evaluate all models from this perspective. In this exhaustive baseline, we comprehensively generate all possible token-level spans of any length from the context document, and systematically compare the query phrase against each target span. However, given that this approach is quadratic with respect to the length of the context document (with the longest context documents in our dataset reaching to more than 3,500 words), we scaled down the experiment to be able to evaluate also using more computationally-intensive embedding-based approaches in addition to lexical methods. We limited our evaluation to a random sample of 1,000 test examples and restricted the maximum span to 100 tokens for the baseline methods, i.e. comparing the query phrase against each context span with a maximum length of 100 tokens. Nonetheless, all ground truth predictions in this evaluation sample are shorter than 100 tokens, therefore the maximum span limitation does not negatively affect the theoretical upper bound (oracle) of our baselines.

The results are shown in Table 4. Our primary span-detection model notably outperforms all baseline models, even when the baselines are evaluated using this exhaustive span-level setting. The span-detection model trained on back-translation data ranks second in both EM and F-score, also surpassing all baselines. All baseline models largely fall short in the EM metric, indicating their difficulty in precisely identifying the target span. In terms of F-score, the drop in performance is less pronounced. However, all baselines clearly underperform when contrasting to the main results where the baselines were limited to sentence-level retrieval. Considering these results, and factoring in the quadratic complexity of the exhaustive all spans approach, this approach seems infeasible for the paraphrase span-detection tasks.

Table 4 A comparison of the Paraphrase-SD model and baseline methods, when evaluating the baselines by exhaustively generating all possible token-level spans from the context documents, and systematically comparing the query phrase against each span

Model	Setup 1	
	EM	F-score
Baselines (all possible spans)		
BM25	8.6	49.1
TF-IDF	11.5	55.4
BERT	30.9	78.0
mSBERT	18.7	75.7
fiSBERT	29.1	77.9
Oracle	100.0	100.0
Paraphrase-SD		
Back-translation	70.3	85.0
Main model	89.2	94.8

6 Error analysis

Next we perform several analyses on the development data in order to better understand the capabilities of our main model and the reasons behind incorrect predictions. Firstly, we automatically categorize the incorrect predictions into several subgroups and inspect the different error groups. Secondly, we calculate the prediction accuracy against the estimated paraphrase complexity in order to investigate whether certain paraphrase categories in the data include more incorrect predictions than others. These experiments are carried out using the development section of the Turku Paraphrase Corpus. Finally, we test the out-of-domain generalization of the model by dividing the paraphrase data into two distinct domains. As the out-of-domain analysis does not require any manual inspection, it is carried out using the test section of the corpus, the numbers then being comparable with the main experiments.

6.1 Error categorization

The incorrect predictions as determined in terms of the exact match are categorized into several subgroups: (1) the model gave an empty prediction even if a valid target exists in the document (false null prediction), (2) the model predicted a span matching one of the negative examples in the corpus, (3) the model predicted a span partially overlapping with the target, further divided into three subgroups: (3.1) the prediction is a substring of the gold segment, (3.2) the gold segment is a substring of the prediction, (3.3) other partial overlap in predicted and gold segments, and (4) other, including cases where the model predicts a segment not overlapping with the gold annotation. The distribution of mispredictions categorized into subgroups is given in Table 5.

While the errors categorized as subgroup (1) and (2) can be seen as clear mispredictions, where the model is not able to identify a paraphrase even if one exists in

Table 5 The error categories of incorrect predictions on the development data

Misprediction type	Setup 1	Setup 2
(1) Null prediction	–	36.50%
(2) Pred not-paraphrase	–	7.17%
(3) Partially correct pred	58.63%	38.67%
(3.1) pred substr. of gold	35.52%	23.86%
(3.2) gold substr. of pred	22.29%	14.16%
(3.3) other partial overlap	0.81%	0.65%
(4) Other	41.37%	17.67%

the document, or identifies the segment annotated as not-a-paraphrase in the original data, the errors belonging to the subgroup (3) contain cases of partially correct predictions where the model is able to identify approximately the correct area from the document, however, the predicted start and end positions slightly differ from the gold segment. From the finer subcategorization it can be seen that in these cases the model is more likely to exclude some part of the gold segment rather than include an additional part. The number of other partial overlaps is negligible. On the other hand, mispredictions in the subgroup (4) include cases which require further manual evaluation. In these cases the model suggests a paraphrase candidate the annotators have not extracted during the corpus construction. These predictions cannot be directly determined to be incorrect, as it is possible that the document includes another occurrence of a paraphrase of the query, which the model then extracts. For many common generic phrases, the probability of more than one correct paraphrase existing in the document is not negligible. Therefore, the evaluation can be considered to give only the lower bound slightly underestimating the actual performance, and we perform a further manual evaluation for the category (4) predictions.

We sample 200 incorrect predictions from the subcategory (4) for the Setup 1 model, and manually annotate for each example whether the predicted span is a valid paraphrase for the query phrase. We find that full 36% of these are in fact valid paraphrases of the query although not being the gold target segment, mostly due to short repeating lines and generic phrases in the movie subtitle section of the corpus, or repeating material between the title, the lead paragraph and the article body in the news article section of the corpus.

6.2 Paraphrase complexity

The paraphrase corpus classifies each paraphrase into one of several classes: *Context dependent* are mutual paraphrases in their present context but not necessarily in other contexts, *context independent* are perfect mutual paraphrases in all reasonably imaginable contexts. In between these two categories, there are near-perfect context independent paraphrases up to one or more qualifying flags: *style* for tone or register difference, *minor difference* marking easily traceable grammatical differences such as person and number, and *subsumption* marking a degree of directionality in the relation, with for instance one mentioning *a woman* while the other *a person*. See

Table 6 Prediction accuracy in Setup 1 for the different paraphrase types annotated in the development set

Paraphrase type	Acc	Support
Context independent	95.0	3,898
with minor diff	92.5	890
with style diff	90.2	902
with subsumption	89.5	8,372
Context dependent	82.0	4,632
Overall	90.2	17,702

Table 7 Prediction accuracy in Setup 1 on several categories of trivial paraphrases in the development set

Category	Acc	Support
Trivial	96	516
same lemmas	91	32
same content word lemmas	97	30
synonym replacement	98	322
content word lemmas with synonym replacement	96	132
Non-trivial	90	17,186
Overall	90	17,702

the annotation guidelines (Kanerva et al., 2021) for more detailed descriptions and examples of these classes. This classification allows us to inspect the model performance w.r.t. the type of the paraphrase, and its degree of context dependence. As seen in Table 6, indeed the Setup 1 model performance clearly correlates with the "degree of universality" of the paraphrases, with 13pp difference between perfect universal paraphrases and context dependent paraphrases.

Given the hypothesis of simpler paraphrases resulting in more confident predictions, Table 7 shows the prediction accuracy across automatically classified "trivial" paraphrase categories following Chang et al. (2021). Here, paraphrases are considered trivial if all their differences can be accounted for with simple, automatically recognizable transformations. These categories include phrases which share the same lemmas thus differing only in word order or inflections, have the same lemmas in terms of content-bearing words only, or if their only differences can be accounted for with a synonym list, or a combination of these. Results show that non-trivial paraphrases, which account for most of the data, more often lead to incorrect predictions compared with trivial paraphrases. However, given the small frequency of trivial paraphrases in the data, the results may suffer from sampling bias.

6.3 Out-of-domain experiments

The majority of the Turku Paraphrase Corpus data is collected from subtitling, only 14% of the pairs being from other domains (e.g. news or discussion forum). To find out how the imbalance of the training data domains affects the prediction ability

we evaluated the main model (Setup 1) separately on evaluation data divided into two parts, subtitle and non-subtitle. As expected, the main model performs better on subtitle data, giving exact match scores of 91.72 for subtitle and 66.06 for non-subtitle. For out-of-domain generalization experiments, we also trained a model using only subtitling data from the training set, and evaluated in the same two domains in order to compare the effect of the small amount of in-domain data in the main model when evaluated on non-subtitle domain. Compared to the main model, this model is only slightly worse on in-domain subtitle data (− 0.1pp EM) likely due to the small decrease in the size of training dataset, however, the decrease is notable on the out-of-domain data (− 4.9pp EM).

7 Discussion and conclusions

We have taken a novel approach towards semantic search by introducing a task of extractive span detection of paraphrases, where given a query phrase, the task is to identify its paraphrased target span from a given document. The primary advantage of this approach is its ability to retrieve a segment of any length, not just a predefined units such as sentences, as is the case with the standard paraphrase retrieval methods utilizing e.g. embedding similarities. Our span detection model trained on the manually annotated Turku Paraphrase Corpus clearly outperformed the retrieval baselines relying on lexical similarity or sentence embeddings by more than 20pp in terms of exact match and 10pp in terms of token-level F-score, demonstrating a clear advantage of the new modelling approach in applications where the retrieval targets are not guaranteed to be sentences. Furthermore, when limiting the evaluation to sentence-level targets only, and therefore eliminating the disadvantage of the sentence-level retrieval baselines, our span-detection model still outperformed the baselines by more than 4pp in terms of EM, and almost 6pp in terms of F-score. We also demonstrated that expanding sentence-level baselines to accommodate non-sentential segments —by exhaustively generating all possible spans of any length— was not a feasible approach for paraphrase-span detection. This was especially evident when considering its retrieval accuracy and the associated quadratic complexity.

Additionally, we have introduced a method for creating artificial paraphrase data through back-translation, suitable for languages where similar paraphrase data including document context is not available. While not achieving the performance of the model trained on the manual paraphrase data, the back-translation model clearly outperforms the sentence embedding baselines in evaluation where the retrieval targets are not guaranteed to be sentences.

Overall, our results indicate that the span detection model is a stronger model of the task as compared to the currently prevailing sentence-based models. In practice, for large-scale extraction purposes, a combination of sentence-based pre-filtering followed by span-based extraction will be desirable so as to manage the computational cost. This would be in line with the practical implementation of other span-based tasks such as question answering. We see pursuing such a method combination as well as expanding the methodology towards a large-scale paraphrase

extraction and semantic search effort a natural continuation of the work presented in this paper.

The source code for all our experiments is available at https://github.com/TurkuNLP/paraphrase-span-detection.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). The work described in this article has received funding from the Academy of Finland and the EU project European Language Grid as one of its pilot projects. The ELG project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 825627.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Agirre, E., Banea, C., Cer, D., et al. (2016). SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, San Diego, California* (pp. 497–511). https://doi.org/10.18653/v1/S16-1081, https://aclanthology.org/S16-1081

Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics* (pp. 4623–4637). https://doi.org/10.18653/v1/2020.acl-main.421, https://aclanthology.org/2020.acl-main.421

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada* (pp. 1–14). https://doi.org/10.18653/v1/S17-2001, https://aclanthology.org/S17-2001.

Chang, L. H., Pyysalo, S., Kanerva, J., & Ginter, F. (2021). Quantitative evaluation of alternative translations in a corpus of highly dissimilar Finnish paraphrases. In *Proceedings of the NoDaLiDa'21 Workshop on Modelling Translation*.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers, pp. 1870–1879). Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-1171, https://aclanthology.org/P17-1171.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics, 8*, 454–470.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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* (Vol. 1, Long and Short Papers, pp. 4171–4186). Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

Dolan, WB., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing IWP*.

Gold, D., Kovatchev, V., & Zesch, T. (2019). Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop. Association for Computational Linguistics, Florence, Italy* (pp. 26–36). https://doi.org/10.18653/v1/W19-4004, https://aclanthology.org/W19-4004

He, Y., Wang, Z., Zhang, Y., Huang, R., & Caverlee, J. (2020). PARADE: A new dataset for paraphrase identification requiring computer science domain knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7572–7582).

Kanerva, J., Ginter, F., Chang, L. H., Rastas, I., Skantsi, V., Kilpelainen, J., & Tarkka, O. (2021). Annotation guidelines for the Turku Paraphrase Corpus. Tech. rep., University of Turku, arXiv:2108.07499.

Kanerva, J., Ginter, F., Chang, L. H., Rastas, I., Skantsi, V., Kilpeläinen, J., Kupari, H. M., Piirto, A., Saarni, J., Sevón, M., Tarkka, O. (2023). Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for finnish. *Natural Language Engineering*. https://doi.org/10.1017/S1351324923000086

Longpre, S., Lu, Y., & Daiber, J. (2020). Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. arXiv preprint arXiv:2007.15207.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., & Wang, H. (2021). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online (pp. 5835–5847). https://doi.org/10.18653/v1/2021.naacl-main.466, https://aclanthology.org/2021.naacl-main.466

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*. Association for Computational Linguistics (pp. 2383–2392). https://www.aclweb.org/anthology/D16-1264.pdf

Rajpurkar, P., Jia, R., Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 2: Short Papers, pp. 784–789). Association for Computational Linguistics, Melbourne, Australia. https://doi.org/10.18653/v1/P18-2124, https://aclanthology.org/P18-2124

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)* (pp. 3982–3992). Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers, pp. 86–96). Association for Computational Linguistics, Berlin, Germany. https://doi.org/10.18653/v1/P16-1009, https://aclanthology.org/P16-1009

Sido, J., Seják, M., Pražák, O., et al. (2021). Czech news dataset for semantic textual similarity. arXiv preprint arXiv:2108.08708.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., & Liu, W. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT - Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2018). Diverse Beam Search: Decoding diverse solutions from neural sequence models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

Vrbanec, T., & Meštrović, A. (2020). Corpus-based paraphrase detection experiments and review. *Information*. https://doi.org/10.3390/info11050241

Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Zeng, C., Li, S., Li, Q., Hu, J., & Hu, J. (2020). A survey on machine reading comprehension-Tasks, evaluation metrics and benchmark datasets. *Applied Sciences, 10*(21), 7640.