*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6841–6851
Marseille, 11–16 May 2020

# Comparative Study of Sentence Embeddings for Contextual Paraphrasing

**Louisa Pragst[1], Wolfgang Minker[1], Stefan Ultes[2]**
[1]Ulm University, Ulm, Germany, {louisa.pragst, wolfgang.minker}@uni-ulm.de
[2]Mercedes-Benz Research & Development, Sindelfingen, Germany, stefan.ultes@daimler.com

## Abstract

Paraphrasing is an important aspect of natural-language generation that can produce more variety in the way specific content is presented. Traditionally, paraphrasing has been focused on finding different words that convey the same meaning. However, in human-human interaction, we regularly express our intention with phrases that are vastly different regarding both word content and syntactic structure. Instead of exchanging only individual words, the complete surface realisation of a sentences is altered while still preserving its meaning and function in a conversation. This kind of contextual paraphrasing did not yet receive a lot of attention from the scientific community despite its potential for the creation of more varied dialogues. In this work, we evaluate several existing approaches to sentence encoding with regard to their ability to capture such context-dependent paraphrasing. To this end, we define a paraphrase classification task that incorporates contextual paraphrases, perform dialogue act clustering, and determine the performance of the sentence embeddings in a sentence swapping task.

**Keywords:** Sentence Similarity, Sentence Embeddings, Sentence Encoding, Paraphrasing

## 1. Introduction

Humans show a lot of variance in the way they express themselves in a conversation. Not only do they change their phrasing by exchanging words with similar meaning for each other, often whole sentences are used interchangeably although they do not have much in common on a surface level. For example, 'When will you be home?' can be answered by 'At six.', 'At six o'clock.' or 'Around six.'. Those variants can be considered word-level paraphrases. However, the sentence 'I just got on the bus.' can fulfil the same function of providing the expected time of arrival, while differing greatly with regard to syntactic structure and the words used. We refer to such sentences that can fulfil the same function in the context of a conversation despite being dissimilar in their surface realisation as *contextual paraphrases*.

In the area of dialogue systems, a number of contributions are concerned with the generation of variety in the utterances of the dialogue system (e.g (Wen et al., 2015; Kozlowski et al., 2003; Langkilde and Knight, 1998)). However, those efforts are mainly focused on word-level paraphrases and little work has been dedicated to the generation of contextual paraphrases. Pragst and Ultes (2018) have made a first effort in this direction by proposing an approach for exchanging sentences, using a dialogue vector model to assess whether two sentences can be used interchangeably. They find that the ability to identify contextual paraphrases is of great importance to the overall performance of the approach. However, the model used in that work is highly specialised to the evaluation corpus and unlikely to perform as well in different scenarios.

This work examines existing approaches to measuring sentence similarity and determines how well they capture information relevant to contextual paraphrasing tasks in a more general setting. To this end, we evaluate four models: sentence similarity based on semantic nets and corpus statistics (Li et al., 2006), BERT (Devlin et al., 2018), skip-thought vectors (Kiros et al., 2015) and InferSent (Conneau et al., 2017). We test their performance regarding paraphrase classification, dialogue act clustering and sentence swapping, and provide an in depth discussion of the implications of our findings.

In the following, we first discuss related work in Section 2. Section 3 gives an overview of the chosen sentence embedding models, followed by a description of our evaluation approaches and discussion of our findings in Section 4. Finally, we summarise our contribution and outline future work.

## 2. Related Work

The evaluation of sentence embeddings is often performed at the time of their introduction with regard to the current state-of-the-art. With this approach, a comparison of different models cannot be found in a single work but must be gathered from numerous individual contributions. Furthermore, the same procedure is not necessarily used across evaluations, further impeding a thorough understanding of advantages and disadvantages of any given model. There have been some efforts to address this issue, e.g. (White et al., 2015). Here, the authors evaluate sentence embeddings with a semantic classification task in a generalised manner. Additionally, the RepEval 2017 Shared Task (Nangia et al., 2017) compares the performance of seven sentence embedding approaches (Chen et al., 2017; Nie and Bansal, 2017; Balazs et al., 2017; Vu et al., 2017; Yang et al., 2017) on a shared task. Those works operate on a strict definition of sentence paraphrases: each sentence must entail the other to be considered a paraphrase. The first task of SemEval-2014 (Marelli et al., 2014) expands on this definition by adding a semantic relatedness score to sentence pairs. The task was solved by 21 participating teams, with 17 submission for the semantic relatedness subtask and 18 for the entailment subtask. However, only 14 of those entries were accompanied by papers (Alves et al., 2014; Beltagy et al., 2014; Bestgen, 2014; Biçici and Way, 2014; Bjerva et al., 2014; Ferrone and Zanzotto, 2014; Gupta et al., 2014; Jimenez et al., 2014; Lai and Hockenmaier, 2014; León et al., 2014;

Lien and Kouylekov, 2014; Proisl et al., 2014; Vo et al., 2014; Zhao et al., 2014).

Often, evaluations of the semantic meaning of sentence embeddings are carried out as classification task on paraphrase corpora such as the MSR paraphrase corpus (Dolan et al., 2004). Another, similar option is the natural language entailment task where, instead of labelling sentence pairs as either paraphrases or unrelated, they are labelled as either entailment, neutral or contradiction. Such corpora include the Stanford Natural Language Inference corpus (Bowman et al., 2015) and the Multi-Genre NLI corpus (Williams et al., 2018), which has been used in the RepEval 2017 Shared Task. The SICK corpus (Marelli et al., 2014) developed for SemEval-2014 Task 1 expands entailment annotations by more fine-grained, human-annotated semantic relatedness scores. In our work, we include a greater number of sentence pairs in our understanding of paraphrases: two sentences are considered contextual paraphrases if they can fulfil the same function in the context of a conversation.

## 3. Sentence Similarity Models

Several approaches to sentence encoding exist, however, not all of them are equally promising to perform well on contextual paraphrasing tasks. Many focus mainly on word-level features such as lexical similarity or word order (e.g. (Sutskever et al., 2014; Palangi et al., 2016; Tsunoo et al., 2017; Shen et al., 2014; Kalchbrenner et al., 2014; Hu et al., 2014; Socher et al., 2011)). The identification of contextual paraphrases is likely to be dependent mainly on the ability of a model to capture functional similarity. Therefore, we choose the following four approaches for our comparative study: a similarity measure based on semantic nets and corpus statistics (Li et al., 2006), BERT (Devlin et al., 2018), skip-thought vectors (Kiros et al., 2015) and InferSent (Conneau et al., 2017). Those models do not rely solely on word similarities for their encoding, but take context into account. In the following, a short overview of the chosen approaches as well as a discussion of their characteristics is given.

### 3.1. Sentence Similarity based on Semantic Nets and Corpus Statistics

The work of Li et al. (2006) introduces a sentence similarity measure that is based on semantic nets and corpus statistics (SNCS). This approach generates sentence embeddings based on a comparison of the words and their relative positions in two sentences and uses those embeddings to estimate a sentence similarity.

SNCS generates two sentence embeddings: a semantic vector that captures similarities between the words of the two embedded sentences, and a word order vector that represents similarities in the word order. Each word found in either of the embedded sentences has a corresponding entry in the embedding vectors. For the semantic embedding, the value of an entry represents how similar the corresponding word is to other words in the sentence: if the word is part of the embedded sentence the entry is set to 1. Otherwise, a similarity score between the word and each word of the embedded sentence is determined based on their distance in a Semantic Net (such as WordNet (Miller, 1995)). The entry is then set to the maximal similarity score. For the word order embedding, the corresponding entry of a word contains the position in the embedded sentence of either the word itself or the most similar word.

The generated sentence embeddings strongly rely on the comparison of the two embedded sentences. They are only valid for a specific sentence pair and change if one of the sentences is replaced. Therefore, this method is not suitable to create general sentence representations, e.g. as input in machine learning models. However, the sentence similarity score that is based on those vectors can be utilised for the decision whether two sentences are contextual paraphrases. It is derived from the cosine distance between the semantic sentence vectors as well as the normalised difference of the word order vectors.

More details regarding the implementation of this sentence similarity measure can be found in the work by Li et al. (2006). We introduce a minor adjustment in our implementation of this approach: the sentence similarity score can be derived from either the cosine distance between semantic sentence vectors, as proposed originally, or the euclidean distance between the two vectors.

### 3.2. BERT

BERT (Devlin et al., 2018) stands for Bidirectional Encoder Representations from Transformers and is trained on the BookCorpus (Zhu et al., 2015) and the textual parts of the English Wikipedia. The training incorporates both word-level information by predicting a word in a sentence and sentence-level information by predicting the following sentence.

BERT utilises a bidirectional Transformer encoder as described by Vaswani et al. (2017). In addition to the commonly used word embeddings, the input to the encoder incorporates the current position in the sentence as well as a sentence affiliation that signifies which sentence the current word belongs to in case the text to be encoded consists of more than one sentence. The sentence encoder is trained on two tasks: predicting a masked word in a sentence as well as predicting the following sentence. The prediction of a masked word is performed using a probability distribution over a fixed vocabulary. The prediction of the consecutive sentence is implemented as binary decision task. The ground truth for this is generated from the corpora by choosing a sentence and either its subsequent sentence or a random sentence.

BERT is intended to be pre-trained in the described manner and then fine-tuned to a specific task. However, as our goal is not to solve a specific task, but rather to determine the informational content of a model, we employ the feature-based approach without fine-tuning presented in the original work. While the results are better when fine-tuning is used, the reported results without it are still promising.

The work of Devlin et al. (2018) describes this approach in more detail. A pre-trained model[1] released by the authors is used in our study.

[1] https://github.com/google-research/bert

### 3.3. Skip-thought Vectors

Skip-thought vectors (Kiros et al., 2015) (STV) are based on the idea that sentences with similar meaning will be used in similar contexts. Therefore, a sentence encoding is learned by predicting surrounding sentences with an encoder-decoder model. The training data consist of continuous text from the BookCorpus dataset (Zhu et al., 2015) that easily allows for the target data to be determined and does not need additional annotation.

The encoder part of the skip-thought model consists of an embedding layer, followed by a GRU (Chung et al., 2014) layer. The words in a sentence are fed sequentially into this model to produce the sentence embedding. The final output of the encoder is used as initial state for the decoder GRU layers during training. Two such layers exist, one for the preceding and one for the subsequent sentence. The input of those layer consists of the sequentially presented embedded words in those sentences, the output is a probability distribution over a fixed vocabulary indicating which word will be next in the target sentence.

A more detailed description of STV can be found in the work of Kiros et al. (2015). The authors furthermore provide a pre-trained model for download[2] that we use in our study.

### 3.4. InferSent

InferSent (InfS) is a supervised approach to learning meaningful sentence encodings taking advantage of natural language inference. It utilises the annotations of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015): 570,000 sentence pairs are manually labelled as either entailment, contradiction or neutral.

The training process encompasses a sentence encoder as well as a classifier. The first step generates sentence embeddings for both sentences in a training pair. Different encoder architectures have been implemented, however, the best results are obtained using a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with max pooling: The embedded words in a sentence are fed sequentially to the biLSTM which generates an output after each time step. The outputs of all time steps are pooled by choosing the maximum value in each dimension. The sentence embeddings, as well as their element-wise product and absolute element-wise difference are then fed into a 3-class classifier. The labels provided in the SNLI corpus serve as target value for this classifier.

The approach is described in more detail in the work of Conneau et al. (2017). We employ the pre-trained sentence encoder[3] provided by the authors.

### 3.5. Comparison of the Approaches

The description of our chosen approaches show that some characteristics are shared while others differ. We discuss the most important differences in this section.

All four considered approaches offer sentence embeddings that can be used as an estimate of how similar two sentences are. However, the embeddings that SNCS provides differ from the other ones significantly in that no generally valid embedding is generated for a single sentence. Instead, the embeddings are dynamically derived from the comparison of two sentences to each other and are only valid in that context. While this excludes them from a number of applications such using them as input to a neural network, it does not interfere with the determination of contextual paraphrases based on the distance between embeddings.

Another aspect that sets SNCS apart from the other approaches is that it is not based on machine learning techniques and therefore does not require training. This distinction impacts several areas of interest: the ability to tailor the embeddings to a specific task, the time needed for training and during deployment and the data needed for each approach.

If pre-trained sentence encoders do not perform in a satisfactory manner, the approaches based on machine learning offer the flexibility to train an encoder on new datasets that may be better suited to the task to try and achieve better results. SNCS can not be improved in that manner.

While neither SNCS nor the machine learning based approaches need time for training if pre-trained encoders are used, the aforementioned flexibility to train a new encoder comes at the cost of additional time needed to perform this training. However, during deployment the machine learning based approaches can embed sentences much faster than SNCS, as they merely perform a number of mathematical operations while SNCS needs to search for words in a Semantic Net repeatedly for each sentence pair individually.

Pre-trained encoders of the machine learning based approaches are the most desirable option with regards to the required data, as they do not need any additional data. In order to train a new encoder, both BERT and STV only require a large amount of continuous text. InfS however needs annotated data and is therefore the most difficult option if a specialised encoder is desired. During deployment, none of the machine learning based approaches needs additional data. However, a semantic net needs to be available for SNCS.

The most consequential characteristic of the considered approaches for contextual paraphrasing tasks is the way in which context is captured within the sentence embeddings. Each approach does this in a unique manner: SNCS relies on the information about words and their connections present in a semantic net. Both BERT and STV take the surrounding sentences into account when training their encoder: BERT embeddings are trained on the decision whether a sentence could follow another one, STV are used to generate both the preceding and the following sentence. InfS encodes the knowledge of human annotators regarding natural language inference.

## 4. Contextual Paraphrasing Tasks

Our comparative study is performed with three contextual paraphrasing tasks: paraphrase classification, dialogue act clustering and sentence swapping. We consider the model that yields the best performance consistently across varying test conditions to be the one currently best suited for contextual paraphrasing tasks. Therefore, we perform our

[2] https://github.com/ryankiros/skip-thoughts

[3] https://github.com/facebookresearch/InferSent

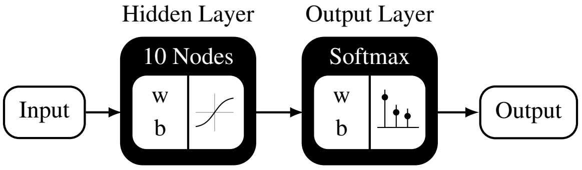

Figure 1: Architecture of the Feedforward Neural Network used to classify paraphrases.

evaluation using not only several tasks, but also several corpora, classifier and distance metrics to ensure a stable performance under varying conditions. As we are more interested in the degree to which the different models capture the context and information relevant for solving general contextual paraphrasing tasks than in optimally solving the presented tasks, we employ simple distance metrics between sentence embeddings as input to our evaluations and abstain from additional and more elaborate inputs such as additional context that might improve the overall results. In the following, this evaluation and our findings are described in more detail.

## 4.1. Contextual Paraphrase Classification

This section is concerned with the task of paraphrase classification. In the first part, we describe our approach to the evaluation of this task, while the second part presents our findings.

### 4.1.1. Task Description

The goal of paraphrase classification is to identify two sentences as either paraphrases or unrelated. The ground truth is usually given by a corpus of sentences pairs with corresponding labels.

We perform the evaluation of the paraphrase classification task in two steps: first, a traditional paraphrase classification task is performed on the MSR paraphrase corpus (Dolan et al., 2004). This corpus contains 5,800 sentences pairs from news sources on the web, human-annotated as either paraphrases or unrelated. The results of this first part serve as comparison for the results obtained in the second part, the contextual paraphrase classification task. Here, we utilise the Opusparcus (Creutz, 2018) corpus.

Opusparcus is a paraphrase corpus for six languages, including English. The sentence pairs are extracted from the Opensubtitles2016 corpus (Lison and Tiedemann, 2016), a corpus of movie and TV subtitles. Opusparcus consists of manually annotated development and test sets, as well as a larger automatically ranked training set. The annotations consist of four categories that rate the degree to which two sentences are paraphrases, *'good'*, *'mostly good'*, *'mostly bad'* and *'bad'*. As opposed to a binary classification, this annotation allows for sentences that are contextual paraphrases to be rated accordingly. Most contextual paraphrases can be found in the category *'mostly good'*, for example the pairs 'No different.'/'That 's the same thing.' or 'I think I got it.'/'I'm good.'. We employ the joint English development and test sets in our study, resulting in 3,088 annotated sentence pairs overall and 1138 sentence pairs in the *'mostly good'* category specifically.

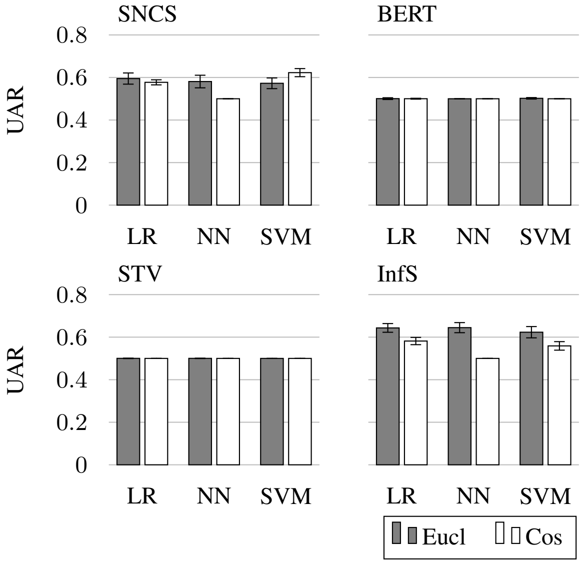

Figure 2: Mean and standard deviation of the UAR achieved for the MSR paraphrase corpus.

| Approach | Accuracy | F-Score |
|---|---|---|
| Vector Based Similarity | 0.65 | 0.75 |
| TF-KLD | 0.80 | 0.86 |
| SNCS | 0.69 | 0.81 |
| BERT | 0.67 | 0.80 |
| STV | 0.67 | 0.80 |
| InfS | 0.71 | 0.81 |

Table 1: Accuracy and F-score for the MSR paraphrase corpus.

Three types of classifiers are considered: Logistic Regression (LR), a Feedforward Neural Network (NN) and a Support Vector Machine (SVM). The architecture of the NN is depicted in Figure 1. It is a simple Multilayer Perceptron with 10 nodes in the hidden layer and a softmax activation function as output. The classifiers are trained and evaluated using ten-fold cross-validation. Their input is composed of either the euclidean or cosine distance between two sentence embeddings.

We utilise the Unweighted Average Recall (UAR) as performance metric, defined as

$$UAR = \frac{1}{N} \sum_{c \in C} \frac{TP_c}{TP_c + FN_c} ,$$

where $C$ is the set of all classes, $N$ the number of classes, $TP_c$ the number of true positives found for class $c$ and $FN_c$ the number of false negatives detected for class $c$. This metric is regarded as balanced even if the classes are of different sizes. Its resulting values range from $0$ to $1$, where $1$ represents a perfect performance.

### 4.1.2. Results of the Study

The traditional paraphrasing task using the MSR paraphrase corpus is best solved by InfS with an UAR of $0.59$.

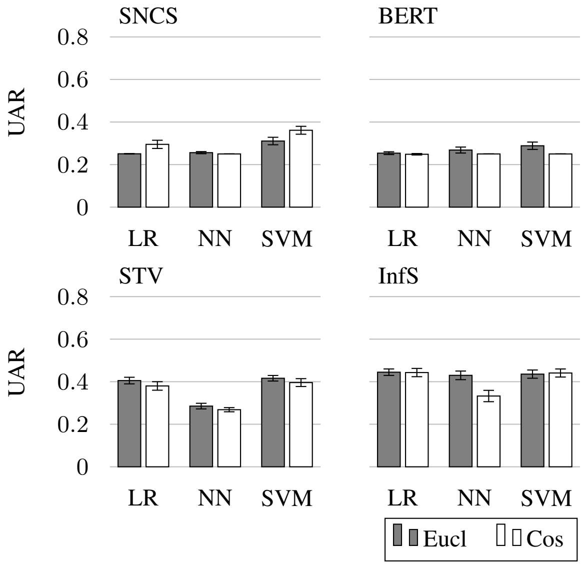

Figure 3: Mean and standard deviation of the UAR achieved for Opusparcus.

With an UAR of 0.57, the SNCS model shows a similar performance that is not significantly different. BERT and STV, however, yield significantly worse results (BERT: $t(59.39) = -13.27$, $p < .001$; STV: $t(59.02) = -13.36$, $p < .001$), as can be seen in Figure 2. With an UAR of 0.5, both do not perform better than the majority class prediction for a binary classification. Presumably, BERT lacks fine-tuning, while STV puts too little focus on the word level to perform well on this task.

We decided to use the UAR for our evaluation as it is robust with regard to unbalanced class sizes. However, many previous works have used Accuracy and F-score as performance metrics. To put our results in context, we provide those values for our chosen approaches in Table 1, as well as the worst and best approaches for Paraphrase Identification according to https://aclweb.org/aclwiki/State_of_the_art (as of 28.02.2020): Vector Based Similarity (Mihalcea et al., 2006) and TF-KLD(Ji and Eisenstein, 2013). Our chosen approaches show moderate success for this task compared to others available.

The evaluation of the contextual paraphrase task results in a different ranking. As Figure 3 shows, the best result is again achieved by InfS. With an average UAR of 0.42, it performs significantly better than the next best model, STV, achieving an UAR of 0.36 ($t(108.24) = 6.36$, $p < .001$).

While InfS can maintain its status, SNCS does not seem to be as well suited for this task, performing significantly worse than STV with an UAR of 0.29 ($t(105.40) = 7.36$, $p < .001$).

A majority class predictor would achieve an UAR of 0.25 for this classification, which both BERT and SNCS barely exceed. However, STV and to an even greater degree InfS surpass this baseline by a rather large margin. Even though there is room for improvement, this provides a solid foundation for further research.

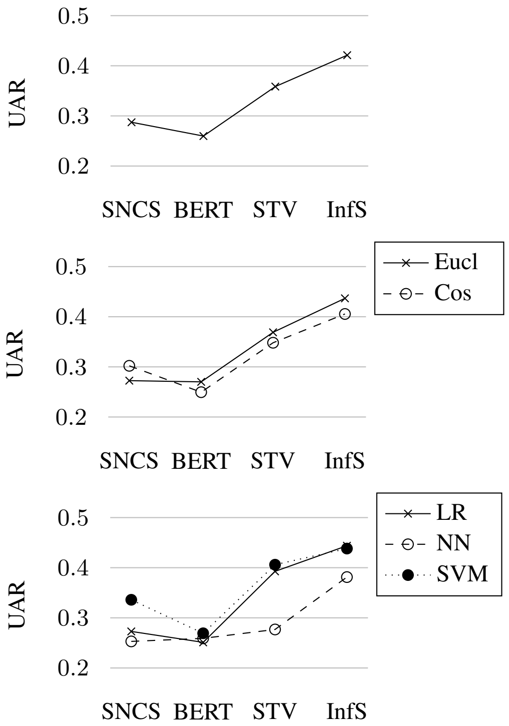

Figure 4: Comparison of sentence similarity models overall, by distance metric and by classifier for Opusparcus. Similar performance trends can be observed across different conditions.

Another interesting finding is that while classifier and distance metric change the maximum UAR that can be achieved, they barely impact the ranking of the different models. The general trend is preserved as can be seen in Figure 4.

## 4.2. Dialogue Act Clustering

In the following, the second part of the comparative study is presented, namely the task of dialogue act clustering. We describe the setup of the study, before assessing our findings.

### 4.2.1. Task Description

Dialogue act clustering determines how well sentences that share a dialogue act are grouped by the sentences encoding. To this end, the clusters given by dialogue act annotations are compared to those found by a clustering algorithm using the distance between two sentence embeddings.

We evaluate this task using the SPAADIA Corpus (Leech and Weisser, 2013) and the Switchboard Dialogue Act Corpus (Jurafsky et al., 1997; Shriberg et al., 1998; Stolcke et al., 2000). Both are dialogue corpora of human-human interaction manually annotated with dialogue acts. The SPAADIA corpus consists of task-oriented dialogues such as train travel booking, whereas the Switchboard corpus contains casual conversations.

The clustering is performed using the k-Means algorithm. As this algorithm is generally implemented to use euclidean distance, we do not consider the cosine distance for this task.

As performance metric, we choose the Adjusted Rand Index (ARI), a variation of the Rand Index corrected for

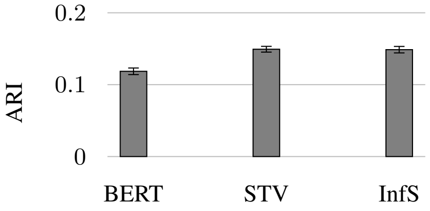

Figure 5: The mean ARI achieved for the clustering of the SPAADIA corpus.

chance. As the UAR for classification, this metric for cluster-similarity is robust for unbalanced cluster sizes. It is defined as follows:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} .$$

The equation encompasses the Rand Index $RI$, the expected Rand Index $E(RI)$ and the maximal Rand Index $max(RI)$:

$$\begin{aligned} RI &= \frac{TP + TN}{TP + FP + TN + FN} , \\ E(RI) &= \frac{(TP + FN)(TP + TN)}{TP + FP + TN + FN} , \\ max(RI) &= \frac{(TP + FN) + (TP + TN)}{2} . \end{aligned}$$

Here, $TP$ is the number of sentence pairs that share a cluster and a dialogue act, $TN$ the number that share neither a cluster nor a dialogue act, $FP$ the number that share a cluster but not a dialogue act, and $FN$ the number that share a dialogue act but not a cluster. The ARI yields values between $-1$ and 1, where 1 is the perfect score.

Over the course of our evaluation, it became apparent that the SNCS model is too time-consuming to allow for a full clustering. As each sentence pair generates a unique encoding, common optimisations that allow for faster computations cannot be employed. Therefore, we exclude this model from our study of the dialogue act clustering task. However, to get an idea of the performance of this model, we add a dialogue act classification task. We generate a ground truth by randomly choosing sentence pairs from the SPAADIA and Switchboard corpus respectively, and annotate whether they share a dialogue act. The generated SPAADIA classification corpus contains 5,464 sentence pairs, the Switchboard classification corpus 6,001. We employ the setup of the paraphrase classification task for this part of the evaluation.

### 4.2.2. Results of the Study

The clustering of the SPAADIA corpus is best solved by STV and InfS with a mean ARI of 0.15. While STV achieves a slightly higher ARI, the difference is not significant. BERT however performs significantly worse than InfS ($t(18) = 14.82$, $p < .001$), as seen in Figure 5. The achieved ARIs are better than the random baseline of 0 but show considerable room for improvement.

The classification part does not completely replicate the ranking: STV performs best with an average UAR of 0.65, significantly outperforming InfS at 0.58 ($t(93.54) = -5.58$, $p < .001$). Both beat the baseline of a majority class prediction by a substantial margin, while SNCS and BERT perform about as well as predicting the majority class would. Figure 6 shows the details of the results.

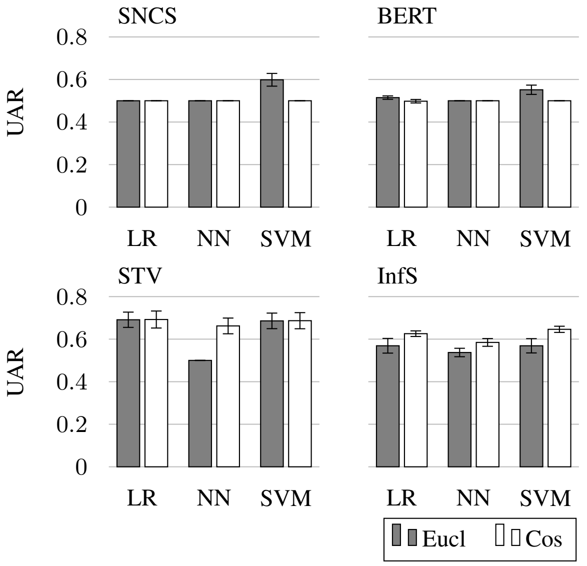

Figure 6: Mean and standard deviation of the UAR achieved for the SPAADIA corpus.

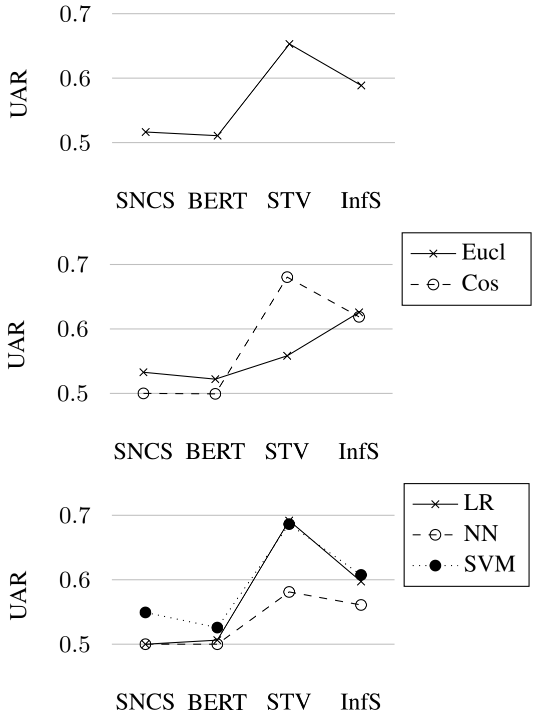

Figure 7: Comparison of sentence similarity models overall, by distance metric and by classifier for the SPAADIA corpus. Similar performance trends can be observed across different conditions.

Again, the general trend of the performance of different models persists across classifiers and distance metrics, as seen in Figure 7.

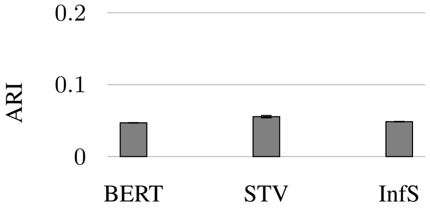

Figure 8: The mean ARI achieved for the clustering of the Switchboard corpus.

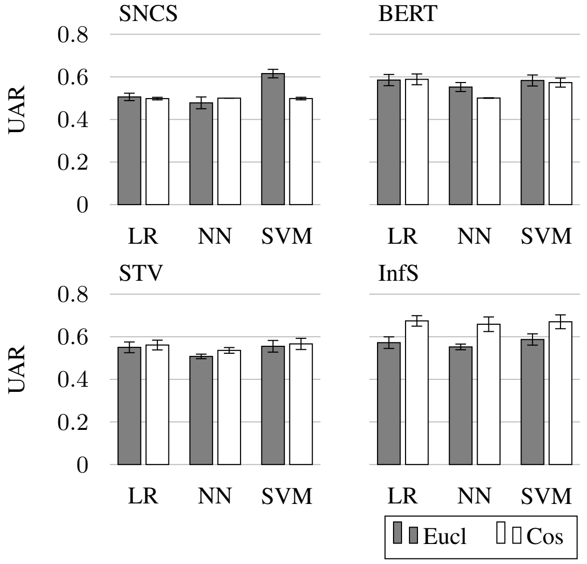

Figure 9: Mean and standard deviation of the UAR achieved for the Switchboard corpus.

Figure 8 shows the mean ARI achieved at the clustering task for Switchboard. Here, STV, with a mean ARI of $0.06$, significantly outperforms the next best model InfS with an ARI of $0.05$ ($t(9.45) = 13.19, p < .001$). This task appears to be more difficult than previous ones: The models barely outperform the random baseline of $0$.

The classification does not reproduce the clustering results. As Figure 9 shows, the best result is achieved by InfS with an UAR $0.62$, which is significantly better than BERT at $0.56$ ($t(102.55) = 6.25$, $p < .001$). This constitutes a substantial improvement from the majority class prediction of $0.5$.

Again, the performance trends share similarities across classifiers and distance metrics, as shown in Figure 10.

### 4.3. Sentence Swapping

In this section, we describe the sentence swapping task and discuss the performance of the different sentence similarity models in this task.

#### 4.3.1. Task Description

The sentence swapping task explores to what degree the chosen sentence similarity models are able to identify sentences that can fulfil the same functionality. The goal is to find an equivalent replacement for the second part of a sentence pair. For example, the second sentence of the sentence pair ‘Do you want to join me for lunch? – I don’t have time today.’ could be replaced by the functionally equivalent ‘I ate a lot already.’. A valid replacement such as this one has to be chosen from a list of available sentences to successfully complete the sentence swapping task.

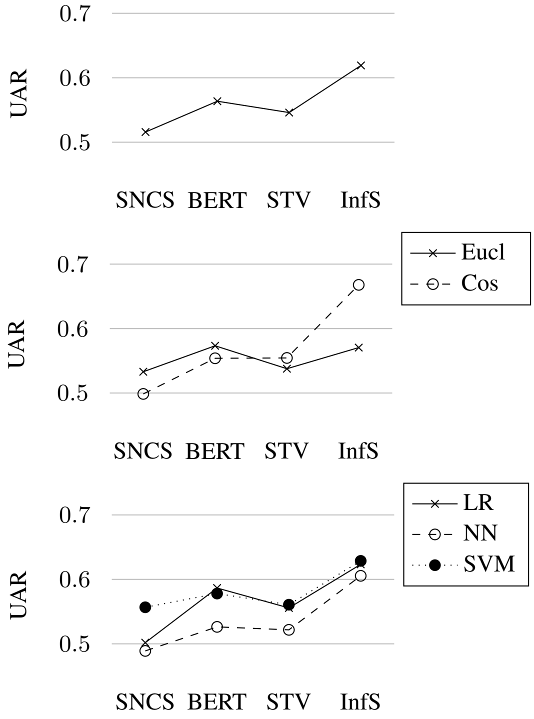

Figure 10: Comparison of sentence similarity models overall, by distance metric and by classifier for the Switchboard corpus. Similar performance trends can be observed across different conditions.

In this part of the evaluation, we replicate the assessment of a dialogue vector model presented by Pragst and Ultes (2018) and adopt it to sentence embeddings that are more generally applicable. The ground truth is given by the automatically generated corpus introduced in that work. This corpus provides sentence pairs in the form of dialogue acts and corresponding verbalisations, e.g. *sa_inviteLunch* – *ua_declineInvitation* verbalised as ‘Do you want to join me for lunch? – I don’t have time today.’. The dialogue acts are extremely fine-grained and describe specific functions such as inviting someone for lunch or declining an invitation. Additionally, the corpus encompasses, for each verbalised sentence, a set of dialogue acts that the sentence can realise. The sentence ‘I ate a lot already.’ can either decline an invitation for lunch as in our previous example or order a small pizza if the first half of the sentence pair was *sa_askSize*, verbalised as ‘What size do you want for your pizza?’. This is reflected in its list of dialogue actions: *ua_declineInvitation* and *ua_sizeSmall*. This list is used to determine whether two sentences can fulfil the same function in a sentence pair: The utterance ‘I don’t have time today.’ shares the dialogue action *ua_declineInvitation* and is therefore a valid exchange candidate in our example sentence pair.

Our evaluation is performed on a randomly chosen set of sentence pairs from the aforementioned dialogue corpus.

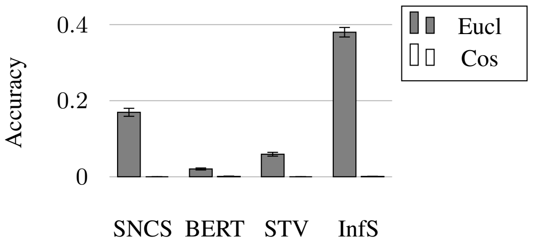

Figure 11: The mean accuracy achieved for the sentence swapping task.

For each sentence pair, the sentence similarity models are used to estimate the similarity score between the second sentence of the pair and all available replacement sentences. The most similar sentence is then chosen for the exchange. The dialogue action corresponding to the second sentence of the pair represents the target dialogue action. The list of dialogue actions that is associated with the replacement sentence has to contain this dialogue action in order for the exchange to be considered successful. The performance metric is accuracy: the percentage of correctly constructed sentence pairs.

#### 4.3.2. Results of the Study

The sentence swapping task is best solved by the InfS model. It achieves a mean accuracy of 0.19 and performs significantly better ($t(26.34) = 2.21$, $p < .05$) than SNCS at 0.09.

The chosen distance metric is of significant importance to this task ($F(1, 72) = 11871.64$, $p < .001$). The cosine distance performs very poorly across all conditions, as can be seen in Figure 11. The best accuracy of 0.38 is achieved by InfS using the euclidean distance.

While Pragst and Ultes (2018) report 0.7 as single best achieved accuracy, their dialogue vector model is highly adapted to the evaluation corpus and unsuitable for other domains. Considering the amount of sentences to choose from, an accuracy of 0.38 can be considered a solid performance for a more generally applicable model such as InfS. Still, further improvements are desirable.

### 4.4. Discussion

In our evaluations, InferSent is consistently among the best performing models. Not only does it achieve good results for the traditional paraphrase classification task on the MSR paraphrase corpus, it is also one of the best models for paraphrase classification on Opusparcus, dialogue act clustering and sentence swapping. Additionally, the chosen classifier and distance metric do not change the ranking of InferSent for most of the tasks. Therefore, it can be considered the best currently available model to solve contextual paraphrasing tasks.

Another interesting finding is that differences between traditional paraphrasing tasks and contextual paraphrasing tasks can be observed: while the sentence similarity based on semantic nets and corpus statistics performs well on the traditional paraphrasing task, skip-thought vectors gain an advantage on contextual paraphrasing tasks. This is most likely due to the semantic net and corpus statistics based similarity measure having a stronger focus on word meaning, while skip-thought vectors are trained taking the context of a sentence into account. BERT is often among the worst performing models. It is likely that the pre-training does not contain enough relevant information to successfully solve the presented tasks and fine-tuning to a specific task is integral to a better performance for this model.

Overall, most of the contextual paraphrasing tasks could be solved with some success over a majority class prediction. The results are comparable to those achieved for traditional paraphrasing tasks. Therefore, existing models provide a solid foundation for research involving contextual paraphrases. However, additional work needs to be done to improve the current results and further advance the handling of contextual paraphrases.

## 5. Conclusion and Future Directions

Traditionally, work on paraphrasing focuses on word-level paraphrases, not taking into account contextual paraphrases: phrases with a common meaning given a specific context. In this work, we assess the ability of four existing approaches to sentence embeddings to rate contextual paraphrases: semantic similarity based on semantic nets and corpus statistics, BERT, skip-thought vectors and InferSent. We assess their performance regarding contextual paraphrase classification, dialogue act clustering and sentence swapping, and find that InferSent achieves a good performance most consistently.

Future work includes further improvement on sentence embedding models for contextual paraphrases, e.g. by advancing or combining the presented models, by finding even more suitable existing models or by creating new approaches for this task. Furthermore, contextual paraphrase generation using those models is a promising field of research that could allow for more variety in the output of dialogue systems.

## 6. Acknowledgements

The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG.

## 7. Bibliographical References

Alves, A., Ferrugento, A., Lourenço, M., and Rodrigues, F. (2014). Asap: Automatic semantic alignment for phrases. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 104–108.

Balazs, J., Marrese-Taylor, E., Loyola, P., and Matsuo, Y. (2017). Refining raw sentence representations for textual entailment recognition via attention. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 51–55, Copenhagen, Denmark, September. Association for Computational Linguistics.

Beltagy, I., Roller, S., Boleda, G., Erk, K., and Mooney, R. (2014). Utexas: Natural language semantics using distributional semantics and probabilistic logic. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 796–801.

Bestgen, Y. (2014). Cecl: a new baseline and a non-compositional approach for the sick benchmark. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 160–165.

Biçici, E. and Way, A. (2014). Rtm-dcu: Referential translation machines for semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 487–496.

Bjerva, J., Bos, J., Van der Goot, R., and Nissim, M. (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40, Copenhagen, Denmark, September. Association for Computational Linguistics.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

Ferrone, L. and Zanzotto, F. M. (2014). half: Comparing a pure cdsm approach with a standard machine learning system for rte. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 300–304.

Gupta, R., Béchara, H., El Maarouf, I., and Orasan, C. (2014). Uow: Nlp techniques developed at the university of wolverhampton for semantic similarity and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 785–789.

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

Hu, B., Lu, Z., Li, H., and Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050.

Ji, Y. and Eisenstein, J. (2013). Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896.

Jimenez, S., Dueñas, G., Baquero, J., and Gelbukh, A. (2014). Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 732–742, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Kozlowski, R., McCoy, K. F., and Vijay-Shanker, K. (2003). Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 1–8. Association for Computational Linguistics.

Lai, A. and Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334.

Langkilde, I. and Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 704–710. Association for Computational Linguistics.

Leech, G. and Weisser, M. (2013). The spaadia annotation scheme. *Retrieved from martinweisser. org/publications/SPAADIA_Annotation_Scheme. pdf (last accessed November 2015)*.

León, S., Vilarino, D., Pinto, D., Tovar, M., and Beltrán, B. (2014). Buap: evaluating compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 145–148.

Li, Y., McLean, D., Bandar, Z. A., Crockett, K., et al. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge & Data Engineering*, (8):1138–1150.

Lien, E. and Kouylekov, M. (2014). Uio-lien: Entailment recognition using minimal recursion semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 699–703, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nangia, N., Williams, A., Lazaridou, A., and Bowman, S. (2017). The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nie, Y. and Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark, September. Association for Computational Linguistics.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., and Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707.

Pragst, L. and Ultes, S. (2018). Changing the level of directness in dialogue using dialogue vector models and recurrent neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 11–19.

Proisl, T., Evert, S., Greiner, P., and Kabashi, B. (2014). Semantiklue: Robust semantic similarity at multiple levels using maximum weight matching. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 532–540, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 101–110. ACM.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):439–487.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., and Van Ess-Dykema, C. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tsunoo, E., Bell, P., and Renals, S. (2017). Hierarchical recurrent neural network for story segmentation. *Proc. Interspeech 2017*, pages 2919–2923.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vo, N. P. A., Popescu, O., and Caselli, T. (2014). Fbk-tr: Svm for semantic relatedeness and corpus patterns for rte. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 289–293, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Vu, H. T., Pham, T.-H., Bai, X., Tanti, M., van der Plas, L., and Gatt, A. (2017). LCT-MALTA's submission to RepEval 2017 shared task. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 56–60, Copenhagen, Denmark, September. Association for Computational Linguistics.

Wen, T.-H., Gasic, M., Mrkšić, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

White, L., Togneri, R., Liu, W., and Bennamoun, M. (2015). How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 9:1–9:8, New York, NY, USA. ACM.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018*

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yang, H., Costa-jussà, M. R., and Fonollosa, J. A. R. (2017). Character-level intra attention network for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 46–50, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhao, J., Zhu, T., and Lan, M. (2014). Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 271–277, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. pages 19–27.