Have We Solved The *Hard* Problem? It's Not *Easy*! Contextual Lexical Contrast as a Means to Probe Neural Coherence

Wenqiang Lei,[1] Yisong Miao,[1] Runpeng Xie,[2] Bonnie Webber,[3]
Meichun Liu,[4] Tat-Seng Chua,[1] Nancy F. Chen[5]
[1]National University of Singapore, [2]Fudan University, [3]University of Edinburgh
[4]City University of Hong Kong, [5]Institute for Infocomm Research, A*STAR
wenqianglei@gmail.com, miaoyisong@gmail.com, runpeng.xie@outlook.com, bonnie@inf.ed.ac.uk, meichliu@cityu.edu.hk,
dcscts@nus.edu.sg, nfychen@i2r.a-star.edu.sg

Abstract

Lexical cohesion is a fundamental mechanism for text which requires a pair of words to be interpreted as a certain type of lexical relation (e.g., similarity) to understand a coherent context; we refer to such relations as the *contextual lexical relation*. However, work on lexical cohesion has not modeled context comprehensively in considering lexical relations due to the lack of linguistic resources. In this paper, we take initial steps to address contextual lexical relations by focusing on the contrast relation, as it is a well-known relation though it is more subtle and relatively less resourced. We present a corpus named $Cont^2Lex$ to make **Cont**extual **Lex**ical **Cont**rast Recognition a computationally feasible task. We benchmark this task with widely-adopted semantic representations; we discover that contextual embeddings (e.g. BERT) generally outperform static embeddings (e.g. Glove), but barely go beyond 70% in accuracy performance. In addition, we find that all embeddings perform better when CLC occurs within the same sentence, suggesting possible limitations of current computational coherence models. Another intriguing discovery is the improvement of BERT in CLC is largely attributed to its modeling of CLC word pairs co-occurring with other word repetitions. Such observations imply that the progress made in lexical coherence modeling remains relatively primitive even for semantic representations such as BERT that have been empowering numerous standard NLP tasks to approach human benchmarks. Through presenting our corpus and benchmark, we attempt to seed initial discussions and endeavors in advancing semantic representations from modeling syntactic and semantic levels to coherence and discourse levels.[1]

1 Introduction

Coherence is what distinguishes well-written text from disorganized text and well-planned dialogue from random sentences and utterances (Halliday and Hasan 2014). It is a fundamental phenomenon in natural languages as long as the texts are consisting of multiple sentences and hence being imperative to downstream tasks such as text quality assessment (Li and Hovy 2014), discourse structure understanding (Somasundaran, Burstein, and Chodorow 2014) and dialogue systems (Lin, Wang, and Lee 1999).

Table 1: We illustrate lexical coherence instantiated by lexical contrast relation. We can see the word pair positive and negative. The **Pos** means this pair establishes a contextual lexical contrast while the **Neg** means not.

Context 1 (Pos.): A positive attitude helps you relax and ace the exams, and a negative mental status will however make you nervous and sleepless.
Context 2 (Neg.): The reviewers are rather positive about this paper. They are nominating it for the Best Paper for its discovery of a negative finding that dispels conventional wisdom.

Therefore, extensive research is focused on coherence modeling. However, efforts have focused on tasks like coreference resolution (Harabagiu and Maiorano 1999) and entity-based solution (*cf* Section 2.1). By contrast, lexical-based approaches are under-explored. According to (Halliday and Hasan 2014), lexical coherence is achieved by a pair of words of certain relations like similar and contrast, being interpreted in a certain context. For example, in Context 1, a reader needs to understand the contrast between "positive" and "negative" so as to interpret the contrasting manners of taking exams – "relax and ace the exam" and "nervous and sleepless", i.e., discourse-level coherence.

Unfortunately, the few existing studies on lexical coherence mostly ignore context. They either directly use lexicons to map the context (Somasundaran, Burstein, and Chodorow 2014) or use word embeddings to model coherence regardless of the context (Mesgar and Strube 2016), which could lead to false positive cases. For example, although the word pair "positive–negative" is recorded as a contrast word pair by ConceptNet (Liu and Singh 2004), it only functions to achieve lexical cohesion in context 1. In fact, in context 2, the two words co-occur just by chance. The words "negative" and "positive" are not interpreted as contrast to understand the coherence of the sentence. Specifically, the "positive" review for the paper is due to an important discovery in the paper, it does not matter whether the result is positive or negative. Hence "negative" is not co-interpreted with "positive" to contribute to the text cohesion of the two sentences in this example.

[1] Corpus and codes are available at: https://cont2lex.github.io/

To facilitate our discussion, in this paper we refer to the lexical relation that is interpreted to function for text cohesion as **contextual lexical relation**. For example, the "positive – negative" pair in *context 1* forms a **contextual lexical contrast** (CLC) relation while it does not in *context 2*.

Although stated to be an important problem grounded in linguistic theory (Halliday and Hasan 2014), few studies directly address contextual lexical relation in a data-driven manner due to the lack of resources. To address this gap, we take the first step by focusing on lexical contrast. We choose contrast because it is one of the most broadly existing types of lexical relation but is reported to be more subtle and less resourced than other lexical relation (e.g. synonymy) (Mohammad, Dorr, and Hirst 2008). We contribute a carefully annotated corpus $Cont^2Lex$ of **Cont**extual **Lex**ical **Cont**rast (CLC) to make the studies on CLC computationally feasible.

Based on this corpus, we carry out a suite of benchmark for the CLC recognition task inclusive of popular static embeddings like GloVe (Pennington, Socher, and Manning 2014) and contextual embeddings like BERT (Devlin et al. 2019) and analyze their behaviors. Our experiments shows that lexical coherence is indeed challenging to model, with the best performing model, BERT, only achieving 70% accuracy on CLC recognition. We further set out to find the characteristics of models. We discover that all models are better at modeling CLC pairs that occur in the same sentence, and the performance is worse when words in the pair are found in different sentences. We also find that the improvement of the strongest model, BERT, can be largely attributed to its modeling of a typical co-occurrence pattern with a pair of repetitive words (Halliday and Hasan 2014) (*cf* Section 5.3). This result suggests contextual modeling still remains at a surface level. These analyses provide a unique angle to get insight of the advancement of semantic representations on lexical coherence modeling, contributing to meaningful discussions and to the recent reflection on the massive success and limitations of these methods (e.g. BERT) (Shwartz and Dagan 2019; Tenney et al. 2019; Manning et al. 2020).

In summary, this paper makes the following contributions:

- We identify a fundamental NLP research problem — contextual lexical contrast (CLC). We contribute a large-scale corpus, $Cont^2Lex$, to make CLC recognition a computational feasible task.
- Our comprehensive benchmark shows that CLC is a challenging task. Further analysis shows that recent advancement is limited to modeling surface textual feature, pushing the development of semantic representations from syntactical and semantic level to coherence and discourse level.

2 Related Work

Being a fundamental phenomena in natural languages, CLC is closely related to three topics in the development of Natural Language Processing. First, it bridges a gap for lexical coherence to consider the context, second it pushes lexical relation into contextual style, and third it provides a unique testbed for existing semantic representations. Therefore, we review those three parts of related works in this section.

2.1 Cohesion Modeling

Text cohesion can be achieved by multiple ways like coreference (Harabagiu and Maiorano 1999), continuity in dialogue(Lei et al. 2018a), discourse connectives (Webber et al. 2019; Lei et al. 2017, 2018b), entity continuities (Barzilay and Lapata 2008) and lexical items (Halliday and Hasan 2014) etc. Among all of them, lexical cohesion is one of the less resourced hence being less studied. In a nutshell, lexical-based cohesion approaches capture two words, each from one clause, that forms certain relation as the signal to glue the two clauses together (Somasundaran, Burstein, and Chodorow 2014; Klebanov and Flor 2013; Mesgar and Strube 2016; Morris and Hirst 2004). A classic piece of work is conducted by Somasundaran, Burstein, and Chodorow (2014) who proposed to model coherence based on lexical chain (Barzilay and Elhadad 1999). However, from early works like (Barzilay and Elhadad 1999), there has been a long standing problem that solutions for lexical cohesion determine the relatedness of words only through the distance between two words and their paths in lexicons (e.g. WordNet) without modeling the context. As discussed in Section 1, the lexical relation here is intrinsically contextually licensed to function for text coherence. One recent work by Mesgar and Strube (2016) is aware of the importance of contextual modeling, but it simply consider sentence as "bag of words". This is partially because there are no annotated resources to distill this problem into a well-defined task to support sophisticated methods. The release of $Cont^2Lex$ corpus addresses this gap, enabling the contextual property to be thoroughly investigated for lexical cohesion.

2.2 Lexical Contrast and Lexical Relations

Computing lexical contrast is a core NLP task which has been studied over years. One traditional way to detect lexical contrast is by fusing various lexical resources (Schwab, Lafourcade, and Prince 2002; Santus et al. 2014; Mohammad et al. 2013). Another type of traditional solution is based on the co-occurrence patterns of word pairs (Justeson and Katz 1991; Fellbaum 1995; Lucerto, Pinto, and Jiménez-Salazar 2002; Roth and Schulte im Walde 2014; Lin et al. 2003). Recently deep learning has empowered us to distinguish multiple lexical relation simultaneously (Glavaš and Vulić 2018; Nguyen and Joty 2017). One limitation of the aforementioned works is that they only focus on building "off-the-shelf" lexicon for the downstream applications without considering the context. Recent works gradually noticed the importance of context modeling for lexical relations. Wang, He, and Zhou (2019) argue that incorporating context is the bottleneck to detect lexical relations. At the same time, a few datasets are also introduced to study lexical semantic such as Stanford Contextual Word Similarity Dataset (SCWS) (Huang et al. 2012) and Word-in-Context (WiC) (Pilehvar and Camacho-Collados 2018) and CoSimLex (Armendariz et al. 2019). They focus on word sense disambiguation problem, hypothesizing different meanings of a word manifested in different context affect lexical relations. However, our $Cont^2Lex$ focus on the text cohesion problem, which is from a totally different perspective. For example, even though the meaning of "positive" and "negative" remains the same in both *Context 1* and *Context 2* (*cf* Section 1), they can behave differently in lexical cohesion.

2.3 Interpretations of Semantic Representations

Semantic representation has been studied for a wide range of tasks, like dialogue (Zhang et al. 2019) and music (Liang et al. 2020). The representation for word is the most fundamental one. From static word embeddings like word2Vec (Mikolov et al. 2013) to contextual word embeddings like BERT (Devlin et al. 2019), more powerful semantic representations have been pushing forward the performance of NLP models with their deeper architectures and bigger model sizes. Despite the good performances, more and more works have calmed down and started to reflect the characteristics of the those representations. Manning et al. (2020) discover that much syntactic information can be captured by self attention pattern of BERT. Tenney et al. (2019) designed a set of "edge-probing" tasks to evaluate BERT's performance on various fundamental semantic and syntactic tasks, and conclude that semantic information is harder for BERT to obtain. Shwartz and Dagan (2019) transform a set of lexical composition tasks into contextual style, and found limited improvement gained by contextual embeddings. Following this line of research, our analyses contributed meaningful discussions on the capability of semantic representation in discourse and coherence modeling. Our corpus can serve as a test bed for further studies on such problems.

3 Corpus

To enable computational approaches to the contextual lexical contrast, we first create an annotated corpus, $Cont^2Lex$, comprising 6,316 instances, with each instance a tuple of (w^+, w^-, c, tag), where w^+ and w^- is a pair of contrast words from existing lexicons. c is the context where w^+ and w^- co-occur, being a sentence or two adjacent sentences. Finally, tag is a binary annotated label, indicating whether contextual lexical contrast holds between w^+ and w^- in c.

3.1 Instance Preparation

To generate candidate instances, we match the contrast word pairs from lexicons with a large number of contexts. To ensure the coherence effect has a higher chance to hold between w^+ and w^-, we constrain our context c to satisfy the minimal spans for text cohesion: (1) w^+ and w^- appear in adjacent sentences; (2) w^+ and w^- appear in the same sentence but different clauses. Note that, to avoid introducing more variables, we control w^+ and w^- to have the same Part-of-speech (POS) tags as recognized by spaCy.

Inspired by the annotation practice in Shwartz and Dagan (2019) and Armendariz et al. (2019), we use the Wikipedia corpus as one source of context to leverage on its broad coverage. We also use the Wall Street Journal Corpus since it is used as a basis to develop many linguistic resources such as PTB (Marcus et al. 1994) and PDTB (Webber et al. 2019), making it easier to extend and expand our work to other NLP tasks. As for the contrast lexicon, we first attempted to apply the lexicon proposed by Mohammad et al. (2013), but unfortunately it turned out to be infeasible to use since more than 90% of the instances do not manifest contextual contrast according to our pilot study. The reason is that the lexicon generated by Mohammad et al. (2013) focuses on coverage, where most of the pairs are only interpreted as contrast in very specific contexts. This high rate (90%) also indicates importance of contextual interpretation: only relying on the lexical relation recorded in lexicons might suffer from high false positive rates in real applications.

After pilot studies, we eventually chose ConceptNet as our lexicon, as it makes a better balance on both precision and coverage. In addition, it also gives a degree score from 0 to 1 for the each contrast pair which we can use to facilitate later annotation. For example, *happy* and *sad* have a score of 1, and *fly* and *walk* have a score of 0.29. We empirically constrained the scores to be above 0.25, since instances below this threshold tend to have much noise as in Mohammad's lexicon. To avoid our corpus being biased to a few frequent word pairs, we limit the maximum appearance of a word pair to 3.

We should note that there might be some collocation pairs that manifest contrast in context, but which are not covered by ConceptNet (e.g., in "Don't try to get fancy, the Wall Street likes modest people.", *fancy* and *modest* are not covered in ConceptNet). For this reason, such pairs are absent from our $Cont^2Lex$ corpus. We leave such pairs for future studies as the annotation is much harder to control.

3.2 Annotation

We recruited five senior undergraduate students majoring in English literature as our on-site annotators . In order to guarantee the annotators fully understand our task, we first gave them a tutorial where we elaborate the definition of contextual lexical contrast. We then administered a quiz on representative cases, and only allowed annotators to continue if they pass the quiz with 100% correctness. We later asked them to do 50 instances of contextual lexical contrast recognition tasks after which we rigorously checked their answers discuss the incorrect cases with the individual annotator. After three such iterations, all the annotators have an at least 90% agreement with us when formal annotations start. During the formal annotation, we insert 5% overlapped instances for monitoring the inter-annotator agreements. The disagreed instances are discussed and used to refine the annotation guidelines. The average speed of annotations is roughly 70 instances per hour.

For each instance, the annotators are given the context c where w^+ and w^- occur, and they are asked to give a binary judgement whether contextual lexical contrast holds between w^+ and w^-. To ensure the annotators paid sufficient attention to the context, we first showed them the whole context but only w^+, then we asked them to find one word in the context that has the most potential to be contrasting with w^+. This enforced the annotators to read the whole context and answer attentively. After they have made a choice, we showed them the real w^- and asked the annotator to conduct real annotation: determining whether w^+ and w^- manifest contrast in c.

Given the intrinsic difficulty of the annotations, we took more steps to ensure the quality of the corpus. Apart from the binary answer options, the annotator can optionally indicate this question is *hard to decide* (HTD), we then examine all the HTD questions and correct their answers when necessary. Our initial inter-annotator agreement(IAA) is 70.6% (this percentage of questions reach full consensus among all 5 annotators). However, as all these HTD questions are further processed by us, our final IAA is calculated by removing those HTD questions, reaching 75.3%. HTD not only helps improve IAA, it relieves the annotators from wasting their time on over-hesitation and accelerates the annotation. Finally our $Cont^2Lex$ Corpus contains 6,316 instances, and 35.7% of them are positive instances that CLC holds[2].

4 Benchmark Method

Our benchmark was chosen to show a model with limited modeling ability on its own, in order to demonstrate the intrinsic power of the embedding methods. We also wanted to investigate if embedding methods (e.g. BERT) have already stored necessary knowledge to recognize CLC, instead of designing fancy model to address CLC.

The benchmark is inspired by recent frameworks including Shwartz and Dagan (2019); Tenney et al. (2019), which systematically incorporate representative embeddings proposed so far, inclusive of both contextual and static word embeddings. In particular, we first convert both the target words and the corresponding context into vector representation and then we encode the input into latent representations, which will be used to conduct the classification task.

4.1 Problem Definition

Given a pair of words w^+ and w^-, we denote their corresponding context as $c = w_1, w_2, ..., w_n$ (note that w^+ and w^- are included in c and we do not specially mark their position). A model needs to predict if contextual contrast holds between w^+ and w^- in the context c. Hence, we define contextual lexical contrast recognition as a binary classification.

4.2 Embedding Methods

We first obtain the word embeddings of each word w_i in context c, resulting in $\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_n$. For static embeddings, we only need to do a simple embedding look-ups for each individual word. For the contextual embeddings, we need to take the whole sentences as inputs to a model to obtain the embeddings. We follow (Shwartz and Dagan 2019) and select the following methods to compare and fix them during training.

Static Word Embeddings:

1. **Glove**: Glove (Pennington, Socher, and Manning 2014) is trained on Wikipedia and Gigawords for word co-occurance preidiction. It has the dimension of 300
2. **Word2Vec**: Word2Vec (Mikolov et al. 2013) is trained on Google News to predict surrounding words given a central word. It also has the dimension of 300.
3. **FastText**: FastText (Joulin et al. 2016) is proposed to refine the above two embeddings methods as it introduces subword embeddings and is suitable for more morphologically-dependent tasks. It was trained for surrounding words predictions on Wikipedia, UMBC and statmt.org, consisting of 300 dimensions.

Contextual Word Embeddings:

1. **ELMo**: ELMo (Peters et al. 2018) is trained on 1B Word Benchmark for a character level language model with deep LSTMs, it has a dimension of 1024.
2. **GPT**: GPT (Radford et al. 2018) is trained on BooksCorpus to obtain a rich language model. It uses transformers instead of LSTM, and encodes BPE's subword instead of word.
   GPT.Lex: We follow Tenney et al. (2019) to set a baseline as "*lexical prior*" abbreviated as GPT.Lex. It uses the context-independent word representation of GPT without any access to surrounding words. We directly take the learned word embedding from GPT model. By doing so, we can evaluate to which extent can GPT utilizes contextual information to aid the prediction of CLC.
3. **BERT**: BERT (Devlin et al. 2019) is also based on transformers, but it is bidirectional as opposed to GPT's single directional architecture. It is trained on two objectives: one is the *Masked Language Model*, which predicts a masked word given the context, another is the *Next Sentence Prediction*. GPT and BERT both have a dimension of 768.
   BERT.Lex: The motivation and operation is identical to the GPT.Lex model.

4.3 Encoders and Classification

After we have obtained the word embeddings $\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_n$, we need to represent them as hidden representations $u_1, u_2, ...u_n$ through encoders for final classification. For encoders, we also follow Shwartz and Dagan (2019):

- **BiLSTM**: BiLSTM has the state-of-the-art performance on many NLP tasks before BERT was introduced. The rationality is that we hope BiLSTM captures additional contextual information for static embeddings. Formally, $u_1, u_2, ...u_n$ = BiLSTM $(\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_n)$.
- **Self Attention**: We concatenate each word's embedding with a weighted sum of other word's embeddings in the context. Formally, $u_i = \mathbf{w}_i \oplus \sum_{j=1}^{n} m_{i,j}\mathbf{w}_j$, where the weight factor m_i is calculated by $\text{Softmax}(\mathbf{w}_i \cdot \mathbf{w})$, which is a inner product with every other word $\mathbf{w}$ followed a softmax. This method does not introduce extra parameters, we hope our complex contextual embeddings can benefit from it.
- **None**: We directly use the embeddings as encoded results. Formally $u_i = \mathbf{w}_i$. The intention behind this is to test if contextual embeddings (e.g. BERT) can achieve superior

[2] Details about our corpus including corpus statistics and sample instances will be uploaded as supplementary material to CMT.

results as advocated in recent literature in a simplist manner. It is worth noting that encoding non-contextual embeddings in this way will totally miss the context information, and we did additional experiment on non-contextual situations.

After we have the encoded representations, we concatenate the u^+ and u^- and use a simple 2-layer MLP for final representation.

5 Experiment

We are driven to find the performance of current semantic representations in CLC and discover their characteristics. Therefore, we direct our experiments with following research questions.

- RQ1: How do models perform on the CLC recognition?
- RQ2: Are models able to recognize lexical contrast out-of-context?
- RQ3: What are the capabilities and limitations of current models?

Dataset: We used the entire $Cont^2Lex$ corpus to conduct experiments, total number of instances is 6,316. We used 5-ply cross validation and randomly split the corpus to partitions of 70%, 15%, and 15% for training, validation and testing.

Implementation details: We used PyTorch to implement our models. All hyper parameters are tuned on the validation set. We used Adam optimizer with the learning rate of 1e-4 and an L2 regularization strength of 1e-5 to prevent overfitting. We used the default hyperparameters for contextual embeddings. We set the hidden size of the BiLSTM as 256 and the hidden size of MLP to 256. Early stopping criteria is performed on the validation set. For each configuration, we conduct experiments for 5-cross validation, and for each validation we run for 5 times with different random seeds and report the average score.

5.1 Main Experiment: CLC Benchmark (RQ1)

We first compare the overall performance of embedding methods regarding different encoders. We also set a **majority baseline** by having all models always predict the majority label, i.e., negative. The results in Table 2 leads to following discoveries:

(1) Though all models outperform the Majority Baseline, the best performing model, BERT only exceeds 70.0 accuracy score, revealing the intrinsic difficulty of detecting contextual lexical contrast.

(2) GPT and BERT significantly outperform their lexical baseline (i.e. GPT.Lex and BERT.Lex) and outperform all static embeddings (i.e. Glove, Word2Vec, FastText), validating their capability of contextual modeling. However, ELMo does not significantly outperform static embeddings, suggesting a possible limitation of its architecture; we leave deeper discussion on this for future work.

Table 2: Main Experiment: We report the accuracy score of contextual lexical contrast recognition by different embedding methods and encoders (RQ1).

	BiLSTM	Attention	None
Glove	65.3	64.9	65.3
Word2Vec	65	65.7	64.7
FastText	66.2	65.5	66.3
ELMo	65.6	65.6	65.7
GPT.Lex	65.8	64.8	64.8
GPT	66.8	67.0	66.9
BERT.Lex	66.4	66.2	66.4
BERT	70.0	69.2	69.1
Majority	64.3		

(3) The encoders (i.e. BiLSTM, Attention) do not necessarily improve performance. As for the static embedding subgroup (e.g. Glove), it may suggest that CLC is too challenging for BiLSTM and Attention to model. As for the contextual embedding subgroup (e.g. GPT), contextual information is already obtained by the contextual embeddings itself, and BiLSTM encoder might not provide additional help. Similar discoveries are also reported in (Shwartz and Dagan 2019), which evaluates lexical composition in context.

5.2 Out-of-context Lexical Contrast Recognition (RQ2)

The experiments above demonstrate the inability of existing semantic representation methods to characterize contextual lexical contrast. However, it is still unclear whether this is due to their insensitivity to out-of-context lexical contrast (i.e., the lexical contrast whose interpretation is not based on context) or due to the difficulty introduced by context. In this section, we design a binary classification experiment to study whether existing semantic representation methods can detect lexical contrast out-of-context.

In our corpus, the word w^+ and w^- in each instance (w^+, w^-, c, tag) is recorded as a pair of out-of-context lexical contrast by ConceptNet regardless of whether they manifest contextual contrast in c. Hence there are naturally positive samples (w^+, w^-) of out-of-context lexical contrast. To conduct a binary classification experiment, we also need to synthesize negative samples of (w^+, $w^{'}$). To make it more comparable with our CLC main experiment, we restrict the negative sampling from the same context c where w^+ and w^- occurs. We further constrain the sampled word $w^{'}$ from c to obtain the same POS as those of w^+ and w^-. We assume that lexical contrast mostly does not hold between $w^{'}$ and w^+ or w^-. Therefore, we choose either (w^+, $w^{'}$) or (w^-, $w^{'}$) as the corresponding negative instances.

We then conduct experiments to ask various models to classify such two classes. Specifically, we use the ”None” encoder for all embeddings, meaning we simply concatenate the embedding and feed them directly to classifiers. In order to make the two sets of experiments as comparable as possible, in our out-of-context lexical contrast recognition

Table 3: Accuracy score of out-of-context lexical contrast recognition (RQ2).

Embedding	Glove	Word2Vec	FastText
Acc.	79.7	82.6	84.1
Embedding	ELMo	GPT	BERT
Acc.	83.5	81.2	79.5

experiment, we treat tuples (w^+, w^-) as positive instances for out-of-context lexical contrast and choose (w^+, $w^{'}$) and (w^-, $w^{'}$) as the corresponding negative instances. We also sample the same ratio of positive instances as contextual lexical contrast experiment. We use the same model design and similar hyper-parameter settings to conduct experiments.

From Table 3 we see that all embedding methods achieve a much higher score than the best performance in Table 2. It is worth noting that since we use ”None” encoding, the static embedding subgroup receives no contextual information, but still obtain substantially higher performance. Therefore we conclude that all embedding methods are able to model lexical contrast to some extent reasonably well when no context is considered. Note that the contextual embedding subgroup (especially GPT, BERT) are slightly inferior to Glove and Word2Vec in this out-of-context contrast recognition task while significantly outperforming the CLC benchmark (as discussed in RQ1). This result is in line with their model design principle that leverage contexts to enhance semantic representation.

5.3 Model Characteristics (RQ3)

In order to gain deeper insight into the model characteristics on contextual lexical contrast, we analyze two typical patterns: (1) whether two words appear in the same sentence but different clauses or from two adjacent sentences; (2) co-occurrence of word repetitions.

CLC Word Pairs Occurring in the Same Sentence: We investigated two types of the contexts: (1) w^+ and w^- appear in the same sentence but different clauses (denoted by S). (2) w^+ and w^- appear in adjacent sentences (as ¬S).

Table 4 shows that contextual modeling is easier for **S**. First, modeling ¬S is challenging, since a number of methods cannot outperform the majority baseline in ¬S. By comparing the improvement from BERT.Lex to the full BERT model, we find significant improvement in S (e.g. 60.7 → 67.4) and very limited in ¬S (e.g. 69.8 → 71.4), suggesting that the contextual modeling capability of BERT mainly lies in the S subcategory.

We also compare BERT and FastText, which are the best performing methods in the contextual and static subgroups. We notice a larger improvement in S (e.g. 60.4 → 67.4) than in ¬S (e.g. 69.8 → 71.4), suggesting the improvement of BERT over static models still lies in the S subcategory. These discoveries also point to possible limitations of BERT in characterizing contextual lexical contrast in different sentences. This finding could be related to the training objective of BERT: it is good at next sentence prediction which focuses on higher level semantic information, but it could gloss over information related to lexical coherence.

Table 4: Model performance on subcategories in S and ¬S. To save space, we only report accuracy scores of the best encoder (i.e. BERT+BiLSTM according to Table 2). (RQ3)

	S	¬S
Glove+None	61.3 (+4.2)	67.9 (-2.0)
W2V+Attention	60.3 (+3.2)	68.8 (-1.1)
FastText+None	60.4 (+3.3)	69.8 (-0.1)
ELMo+None	63.6 (+6.5)	68 (-1.9)
GPT.Lex+BiLSTM	61.5 (+4.4)	68.3 (-1.6)
GPT+Attention	64 (+6.9)	68.7 (-1.2)
BERT.Lex+BiLSTM	60.7 (+3.6)	69.8 (-0.1)
BERT+BiLSTM	67.4 (+10.3)	71.4 (+1.5)
Majority	57.1	69.9

Table 5: Corpus study for word repetitions co-occurring with CLC pairs. Repeated parts are bolded.

(Ex. 1 Repetition) ...is considered spurious **by** Hefele questionable by Haddan and Stubbs, and genuine **by** JaffA Regest.
(Ex. 2 Repetition) They had many children who **lived in the** darkness between them. The children wished to **live in the** light and so separated their unwilling parents.

Table 6: Model performance on subcategories within **R** and without **R**. We only report accuracy scores of the best encoder to save space. (RQ3)

	R	¬R
Glove+None	60.9 (+7.2)	67.3 (-3.1)
W2V+Attention	60.4 (+6.7)	68.1 (-2.3)
FastText+None	61.1 (+7.4)	68.8 (-1.6)
ELMo+None	63 (+9.4)	68 (-2.5)
GPT.Lex+BiLSTM	60.8 (+7.1)	68.1 (-2.3)
GPT+Attention	65.5 (+11.8)	67.8 (-2.6)
BERT.Lex+BiLSTM	58.7 (+5.0)	69.9 (-0.4)
BERT+BiLSTM	68.7 (+14.9)	70.7 (+0.3)
Majority	53.7	70.4

Word Repetitions Co-Occurring with CLC Pairs: We find that a pair of repetitive word(s) tend to co-occur with w^+ and w^- when w^+ and w^- manifest CLC. As Ex. 1 in Table 5, the recognition of positive CLC between “spurious” and “genuine” can be aided by the reoccurring preposition “by”. Similarly, the reoccurring words “live in the” might have made detecting CLC an easier task in Ex. 2.

To investigate this observation, we split the entire test set to two: one consists of instances with repetitive words near w_1 and w_2 in a window size of 3[3] (denoted as **R**);

[3] We have experimented with multiple window sizes but only

the other set consists the remaining examples (denoted as ¬R). As we can see in Table 6, Without the repetition patterns (¬R), the performance of all models can barely outperform the majority baseline, implying the intrinsic difficulty of contextual modeling without such explicit patterns of word repetitions. By comparing inter-model performance in ¬R, we find very limited improvement brought by BERT (the strongest model). Specifically, the improvement from BERT.Lex+BiLSTM to BERT+BiLSTM is 69.9 → 70.7, and that from FastText+None to BERT+BiLSTM is 68.8 → 70.7. However, the improvement for R subcategory is much larger: 58.7 → 68.7 from BERT.Lex+BiLSTM to BERT+BiLSTM and 61.1 → 68.7 from FastText+None to BERT+BiLSTM. These intriguing discoveries suggest the improvements from static embeddings to BERT are largely attributed to modeling simple repetitions.

5.4 Discussion on Other Typical Cohesive Ties

In addition to a pair of repetitive words, there are siblings of this pattern in linguistic coherence theory (Collins 2012; Halliday and Hasan 2014), called *cohesive ties*[4]. Typical cohesive ties that appear in a pair include repetition, coreference, substitution, ellipsis and collocation (Halliday and Hasan 2014). We discover that the co-occurrence pattern discussed before not only applies to word repetition, but also to other typical cohesive ties. For examples in Table 7, in Ex. 3 although "way" and "path" are not repeated words, they are semantically similar, which could still aid the CLC recognition task. In Ex. 4, "one" is used instead of repeating "countries", it may aid the modeling of CLC. Finally, understanding "it" is referred to " Party Control" in Ex. 5 could also help the recognition task.

Can existing representations leverage other cohesive ties for CLC recognition? To gain more insights, we conduct a case study by annotating 200 randomly sample CLC cases in our test set and followed the same experiment steps (*cf* Section 5.1). For each test instance we decide if one of the five cohesive ties (T) co-occur with w^+ and w^-, resulting in two subsets T and ¬T. To make the results comparable with Table 6, we also set the context window to be 3. Note that simple repetition (R) is a subset of cohesive tie (T), and thus ¬R is a superset of ¬T, with more instances from other cohesive ties as elaborated in Table 7. Therefore by comparing the performance between ¬R and ¬T, we can estimate how models are using other cohesive ties for CLC recognition.

Due to space constraints, we only illustrate the representative result — the improvement of BERT, the strongest model, over Glove, the weakest model. Table 8 shows that though ¬R include cohesive ties that were not shown in ¬T, its improvement is nearly the same as ¬T. This finding suggests that BERT may not utilize cohesive ties (other than simple repetition) for lexical coherence modeling. Together with our discovery in last section, we find that our contextual model's improvement is mainly attributed to simple repetitive patterns, thus not fully exploiting the modeling potential of other cohesive ties. We hope that our case study and analysis in this work serves as a springboard to initiate dialogues and endeavors at the interface of contextual lexical contrast and neural coherence modeling.

Table 7: Corpus study for other cohesive ties.

(Ex. 3 Collocation) Theseus decided to go to Athens and had the choice of going by sea, which was the safe **way** or by land, following a dangerous **path** with thieves and bandits all the way.
(Ex. 4 Substitution) Although accepted by many **countries**, his proposal was rejected by **one** in Africa.
(Ex. 5 Coreference) **Party control** is tightest in government offices and in urban economic, industrial, and cultural settings; **it** is considerably looser in the rural areas.

Table 8: The improvement of BERT over the weakest model, Glove. BERT achieves +4.2 improvement in ¬T over Glove (both with BiLSTM encoder). We follow the fashion in main experiment to report accuracy averaged by random seeds.

	¬R	¬T
ΔBERT+BiLSTM	+4.1	+4.2
ΔBERT+Attention	+3.6	+3.5
ΔBERT+None	+3.7	+3.7

6 Conclusion and Future Work

In this paper we study an intriguing phenomena as Contextual Lexical Contrast (CLC). CLC holds between two words if they are interpreted as contrast to understand the coherence of the context. We contributed a well-annotated corpus, $Cont^2Lex$ to make CLC recognition a computationally feasible task. We found that model performance is generally unsatisfactory, even for BERT, which was the best performing model. Analysis results reveal that detecting lexical contrast in out-of-context scenarios is feasible with the applied models, suggesting that the difficulty of CLC likely stems from contextual modeling. In particular, we found that performance gains from BERT are largely attributed to cohesion ties that are manifested in simple repetition form, while other cohesive ties are under-exploited. This finding suggests that even though contextualized language models such as BERT that have been successful in enabling many NLP tasks to reach human benchmarks, this is not the case for contextual lexical contrast recognition. This probing analysis implies that there is still much space for improvement for language representation learning to model discourse and coherence more effectively.

Future works include extending our work to other contextual lexical relations such as synonymy and hypernymy, expanding our scope from lexical items to multi-word expressions, and quantifying how CLC might benefit downstream tasks such as discourse relation recognition. It is also interesting to investigate the possible synergy between modeling lexical-level and sentence-level coherence.

show results where window size is 3, as the results are relatively insensitive to window size.

[4] According to linguistic theories (Collins 2012; Halliday and Hasan 2014), cohesive ties refer to words/phrases that link different pieces of writing (usually two clauses or sentences) together.

Acknowledgements: We thank Prof Min-Yen Kan and Dr Vered Shwartz for valuable discussions with them. We also thank the anonymous reviewers for their valuable comments.

This research is partially supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Armendariz, C. S.; Purver, M.; Ulčar, M.; Pollak, S.; Ljubešić, N.; Granroth-Wilding, M.; and Vaik, K. 2019. CoSimLex: A Resource for Evaluating Graded Word Similarity in Context. *arXiv:1912.05320* .

Barzilay, R.; and Elhadad, M. 1999. Using lexical chains for text summarization. *Advances in automatic text summarization* 111–121.

Barzilay, R.; and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1): 1–34.

Collins, M. 2012. Library Guides: Interactive Rubric for Written Communication: Persuasive Essay. *Genre* 2(2.1): 2–2.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Fellbaum, C. 1995. Co-occurrence and antonymy. *International journal of lexicography* 8(4): 281–303.

Glavaš, G.; and Vulić, I. 2018. Discriminating between Lexico-Semantic Relations with the Specialization Tensor Model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 181–187. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-2029. URL https://www.aclweb.org/anthology/N18-2029.

Halliday, M. A. K.; and Hasan, R. 2014. *Cohesion in english*. Routledge.

Harabagiu, S.; and Maiorano, S. J. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *The Relation of Discourse/Dialogue Structure and Reference*.

Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–882.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. Fasttext. zip: Compressing text classification models. *arXiv:1612.03651* .

Justeson, J. S.; and Katz, S. M. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational linguistics* 17(1): 1–19.

Klebanov, B. B.; and Flor, M. 2013. Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1148–1158.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018a. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1437–1447.

Lei, W.; Wang, X.; Liu, M.; Ilievski, I.; He, X.; and Kan, M.-Y. 2017. SWIM: A Simple Word Interaction Model for Implicit Discourse Relation Recognition. In *IJCAI*, 4026–4032.

Lei, W.; Xiang, Y.; Wang, Y.; Zhong, Q.; Liu, M.; and Kan, M.-Y. 2018b. Linguistic Properties Matter for Implicit Discourse Relation Recognition: Combining Semantic Interaction, Topic Continuity and Attribution. In *AAAI*, 4848–4855.

Li, J.; and Hovy, E. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2039–2048.

Liang, H.; Lei, W.; Chan, P. Y.; Yang, Z.; Sun, M.; and Chua, T.-S. 2020. PiRhDy: Learning Pitch-, Rhythm-, and Dynamics-aware Embeddings for Symbolic Music. In *Proceedings of the 28th ACM International Conference on Multimedia*, 574–582.

Lin, B.-s.; Wang, H.-m.; and Lee, L.-s. 1999. A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history .

Lin, D.; Zhao, S.; Qin, L.; and Zhou, M. 2003. Identifying synonyms among distributionally similar words. In *IJCAI*, volume 3, 1492–1493.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22(4): 211–226.

Lucerto, C.; Pinto, D.; and Jiménez-Salazar, H. 2002. An automatic method to identify antonymy. In *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, 105–111. Puebla.

Manning, C. D.; Clark, K.; Hewitt, J.; Khandelwal, U.; and Levy, O. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences (PNAS)* .

Marcus, M.; Kim, G.; Marcinkiewicz, M. A.; MacIntyre, R.; Bies, A.; Ferguson, M.; Katz, K.; and Schasberger, B. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, 114–119. Association for Computational Linguistics.

Mesgar, M.; and Strube, M. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the*

2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1414–1423.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Mohammad, S.; Dorr, B.; and Hirst, G. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 982–991. Association for Computational Linguistics.

Mohammad, S. M.; Dorr, B. J.; Hirst, G.; and Turney, P. D. 2013. Computing lexical contrast. *Computational Linguistics* 39(3): 555–590.

Morris, J.; and Hirst, G. 2004. Non-classical lexical semantic relations. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, 46–51.

Nguyen, D. T.; and Joty, S. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1320–1330.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv:1802.05365* .

Pilehvar, M. T.; and Camacho-Collados, J. 2018. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121* .

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* .

Roth, M.; and Schulte im Walde, S. 2014. Combining Word Patterns and Discourse Markers for Paradigmatic Relation Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 524–530. Baltimore, Maryland: Association for Computational Linguistics. doi:10.3115/v1/P14-2086. URL https://www.aclweb.org/anthology/P14-2086.

Santus, E.; Lu, Q.; Lenci, A.; and Huang, C.-R. 2014. Taking antonymy mask off in vector space. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, 135–144.

Schwab, D.; Lafourcade, M.; and Prince, V. 2002. Antonymy and conceptual vectors. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7. Association for Computational Linguistics.

Shwartz, V.; and Dagan, I. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics* .

Somasundaran, S.; Burstein, J.; and Chodorow, M. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, 950–961.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Van Durme, B.; Bowman, S. R.; Das, D.; et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316* .

Wang, C.; He, X.; and Zhou, A. 2019. SphereRE: Distinguishing Lexical Relations with Hyperspherical Relation Embeddings. In *ACL*, 1727–1737. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1169. URL https://www.aclweb.org/anthology/P19-1169.

Webber, B.; Prasad, R.; Lee, A.; and Joshi, A. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* .