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

AAAI 2021
Full Slides

Wenqiang Lei, **Yisong Miao**, Runpeng Xie, Bonnie Webber, Meichun Liu, Tat-Seng Chua and Nancy F. Chen

https://cont2lex.github.io/
Table of Contents

1. Task Introduction — Contextual Lexical Contrast
2. Motivation and Background
3. Cont²Lex Corpus
4. Benchmark
5. Experiments and Conclusions
1. Task Introduction — Contextual Lexical Contrast
Contextual Lexical Contrast (CLC)

Example: positive vs negative:

(Ex. 1 Positive CLC): A positive attitude helps you relax and ace the exams, and a negative mental status will however make you nervous and sleepless.

(Ex. 2 Negative CLC): 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.

Definition of CLC (a new NLP task):
- Two words from the same sentence (or adjacent sentences) form a "contextual lexical contrast" word pair if these two words exhibit contrastive semantics that contribute to the coherence of sentential context.
2. Motivation and Background
Motivation and Background
— Why CLC is important.

- Cohesion Modeling
  - Entity-based
  - Lexical-based
- Lexical Contrast and Lexical Relation
- Interpretations of Semantic Representations
Cohesion Modeling
— Lexical-based approach is overlooked.

Entity-based Approach

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) is also based on the transformer, but it is bidirectional as opposed to left-to-right as in the OpenAI GPT, and the directions are dependent as opposed to ELMo’s independently trained left-to-right and right-to-left LSTMs. It also introduces a somewhat different objective called “masked language model”: during training, some tokens are randomly masked, and the objective is to restore them from the context.

Excerpted from Shwartz & Dagan, TACL2019

✅ Entity grid method (Barzilay and Lapata)

Lexical-based Approach

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) is also based on the transformer, but it is bidirectional as opposed to left-to-right as in the OpenAI GPT, and the directions are dependent as opposed to ELMo’s independently trained left-to-right and right-to-left LSTMs. It also introduces a somewhat different objective called “masked language model”: during training, some tokens are randomly masked, and the objective is to restore them from the context.

Excerpted from Shwartz & Dagan, TACL2019

❓ - Being Largely Ignored
- Need to put into context
Lexical Contrast
— Context is critical for downstream applications.

Computing Lexical Contrast

Saif M. Mohammad*
National Research Council Canada

Bonnie J. Dorr**
University of Maryland

Graeme Hirst†
University of Toronto

Peter D. Turney‡
National Research Council Canada

Applications

• Discourse relation.
  “Tokyo is cold. Beijing is hot.”

• Contradiction detection.
  “Kyoto has a predominantly wet climate” / “It is mostly dry in Kyoto”

• Humour detection.
Interpretations of Semantic Representations
— Right timing to do CLC.

- Syntactic tasks: POS, Constituents, Dependencies
- Semantic tasks: SRL, OntoNotes coref, Semantic proto-role

- Light Verb Construction (LVC): make a decision
- Verb-Particle Construction (VPC): carry on vs carry
3. Cont\textsuperscript{2}Lex Corpus
Problem Formalization:
Given $w^+$ and $w^-$ in context $c$ (a sequence of words $w_1, w_2, \ldots w_n$), a human (or a machine) needs to indicate a binary tag for CLC.
Instance Preparation

- Constraint 1: Contrasting degree in ConceptNet
- Constraint 2: Distance between $w^+$ and $w^-$ (Adjacent sentence or difference clause in same sentence.)
- Constraint 3: Appearance of the same pair of $w^+$ and $w^-$

6,316 instances to be annotated.
(Ex. 1 Positive CLC): A **positive** attitude helps you relax and ace the exams, and a **negative** mental status will however make you nervous and sleepless.

(Ex. 2 Negative CLC): 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.

- Quality Control 1: Predict $w^-$, given only $w^+$ and $c$
Corpus Statistics

Inter-Annotator Agreement (IAA):
We calculate IAA using the consensus of our 5 annotators, reaching 75.3%.

<table>
<thead>
<tr>
<th>Part-of-Speech</th>
<th>#</th>
<th>Positive Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>2,413</td>
<td>33.2%</td>
</tr>
<tr>
<td>Verb</td>
<td>1,568</td>
<td>27.9%</td>
</tr>
<tr>
<td>Adj</td>
<td>2,081</td>
<td>43.7%</td>
</tr>
<tr>
<td>Adv</td>
<td>254</td>
<td>40.9%</td>
</tr>
<tr>
<td>Total</td>
<td>6,316</td>
<td>35.7%</td>
</tr>
</tbody>
</table>

Possible reason: Adj and Adv has purer semantic dimensions.
4. Benchmark
- 6,316 instances enable us to do supervised learning, for the binary classification.
- Similar approach as Tenney et.al, and “Embed — Encode — Predict” framework (Shwartz and Dagan)
- We didn’t fine-tune BERT. Why?

Probing Contextual LMs (Tenney et.al. ICLR ’19)
Evaluated Embeddings

- Static embeddings: Glove, Word2Vec, fastText
- Contextual Embeddings: ELMo, OpenAI GPT, BERT
- The “Lex” version of GPT and BERT. Why?

Probing Contextual LMs (Tenney et.al. ICLR ’19)
5. Experiments and Conclusion
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?
### Main Experiment (RQ1)

<table>
<thead>
<tr>
<th>Model</th>
<th>BiLSTM</th>
<th>Attention</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>65.3</td>
<td>64.9</td>
<td>65.3</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>65</td>
<td>65.7</td>
<td>64.7</td>
</tr>
<tr>
<td>FastText</td>
<td>66.2</td>
<td>65.5</td>
<td>66.3</td>
</tr>
<tr>
<td>ELMo</td>
<td>65.6</td>
<td>65.6</td>
<td>65.7</td>
</tr>
<tr>
<td>GPT.Lex</td>
<td>65.8</td>
<td>64.8</td>
<td>64.8</td>
</tr>
<tr>
<td>GPT</td>
<td>66.8</td>
<td>67.0</td>
<td>66.9</td>
</tr>
<tr>
<td>BERT.Lex</td>
<td>66.4</td>
<td>66.2</td>
<td>66.4</td>
</tr>
<tr>
<td>BERT</td>
<td>70.0</td>
<td>69.2</td>
<td>69.1</td>
</tr>
<tr>
<td>Majority</td>
<td></td>
<td>64.3</td>
<td></td>
</tr>
</tbody>
</table>

Acc scores show that CLC is a challenging task!

BERT and GPT are better than their Lex version.
Out-of-context Lexical Contrast (RQ2)

(Ex. 2 Negative CLC): 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.

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>Glove</th>
<th>Word2Vec</th>
<th>fastText</th>
<th>ELMo</th>
<th>GPT</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc.</td>
<td>79.7</td>
<td>82.6</td>
<td>84.1</td>
<td>83.5</td>
<td>81.2</td>
<td>79.5</td>
</tr>
</tbody>
</table>

Acc scores of out-of-context lexical contrast recognition, which is much more higher than CLC.

[Contextual] hard  Lexical [Contrast] easy
Model Characteristics (RQ3)

S: CLC Word Pairs Occurring in the Same Sentence.

R: Word Repetitions Co-Occurring with CLC Pairs.

(Ex. 3 Repetition): ...is considered spurious by Hefele questionable by Haddan and Stubbs, and genuine by JaffA Regest.

(Ex. 4 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.
## Model Characteristics (RQ3)

<table>
<thead>
<tr>
<th>Model Characteristics</th>
<th>S</th>
<th>-S</th>
<th>R</th>
<th>-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove+None</td>
<td>61.3 (+4.2)</td>
<td>67.9 (-2.0)</td>
<td>60.9 (+7.2)</td>
<td>67.3 (-3.1)</td>
</tr>
<tr>
<td>W2V+Attention</td>
<td>60.3 (+3.2)</td>
<td>68.8 (-1.1)</td>
<td>60.4 (+6.7)</td>
<td>68.1 (-2.3)</td>
</tr>
<tr>
<td>FastText+None</td>
<td>60.4 (+3.3)</td>
<td>69.8 (-0.1)</td>
<td>61.1 (+7.4)</td>
<td>68.8 (-1.6)</td>
</tr>
<tr>
<td>ELMo+None</td>
<td>63.6 (+6.5)</td>
<td>68 (-1.9)</td>
<td>63 (+9.4)</td>
<td>68 (-2.5)</td>
</tr>
<tr>
<td>GPT.Lex+BiLSTM</td>
<td>61.5 (+4.4)</td>
<td>68.3 (-1.6)</td>
<td>60.8 (+7.1)</td>
<td>68.1 (-2.3)</td>
</tr>
<tr>
<td>GPT+Attention</td>
<td>64 (+6.9)</td>
<td>68.7 (-1.2)</td>
<td>65.5 (+11.8)</td>
<td>67.8 (-2.6)</td>
</tr>
<tr>
<td>BERT.Lex+BiLSTM</td>
<td>60.7 (+3.6)</td>
<td>69.8 (-0.1)</td>
<td>58.7 (+5.0)</td>
<td>69.9 (-0.4)</td>
</tr>
<tr>
<td>BERT+BiLSTM</td>
<td>67.4 (+10.3)</td>
<td>71.4 (+1.5)</td>
<td>68.7 (+14.9)</td>
<td>70.7 (+0.3)</td>
</tr>
<tr>
<td>Majority</td>
<td>57.1</td>
<td>69.9</td>
<td>53.7</td>
<td>70.4</td>
</tr>
</tbody>
</table>

The delta over baseline are majorly achieved by S and R.
Q: Besides Repetition, what other cohesive ties is BERT using?

Cohesive devices (M.A.K. Halliday):
- Collocation
- Substitution
- Coreference

T: All types of cohesive ties
R: Repetition
R is a subset of T.

This table shows that models are no better handling T than R.
Conclusion

- We propose a new NLP task as CLC for cohesion modelling. Our Cont\textsuperscript{2}Lex corpus makes CLC a computational feasible task.

- CLC is a challenging semantic representation task. Contextual embeddings are capable to capture part of contextual information.

- The advantage gained by BERT is largely due to modelling surface textual patterns.