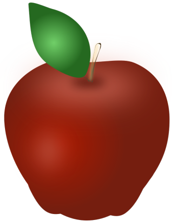
# Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis

Mario Giulianelli, Iris Luden, Raquel Fernández, Andrey Kutuzov

University of Oslo, University of Amsterdam

# Contents

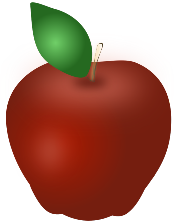
# Contextualized definitions as word representations



## What word representations we use in NLP?

1. Dense embeddings: '*apple*' is `[0.44, 0.32, 0.76 ... 0.01]`
   - Can be learned automatically, convenient in modeling, but not human-readable
2. Word definitions: '*apple*' is 'EDIBLE POME FRUIT OF A USUALLY CULTIVATED TREE OF THE ROSE FAMILY'
   - Human readable and interpretable, but expensive to create and difficult to use in modelling

# Contextualized definitions as word representations

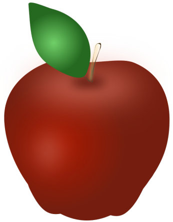


**What word representations we use in NLP?**

1. Dense embeddings: '*apple*' is `[0.44, 0.32, 0.76 ... 0.01]`
   - Can be learned automatically, convenient in modeling, but not human-readable
2. Word definitions: '*apple*' is 'EDIBLE POME FRUIT OF A USUALLY CULTIVATED TREE OF THE ROSE FAMILY'
   - Human readable and interpretable, but expensive to create and difficult to use in modelling

What if we could have the best of both worlds?

# Contextualized definitions as word representations



## What word representations we use in NLP?

1. Dense embeddings: '*apple*' is `[0.44, 0.32, 0.76 ... 0.01]`
   - Can be learned automatically, convenient in modeling, but not human-readable
2. Word definitions: '*apple*' is 'EDIBLE POME FRUIT OF A USUALLY CULTIVATED TREE OF THE ROSE FAMILY'
   - Human readable and interpretable, but expensive to create and difficult to use in modelling

What if we could have the best of both worlds?

## Definition generated by a fine-tuned language model

| | | |
|---|---|---|
| 'I frequently saw Mehevi and several other **chefs** and warriors of note take part' | **chef** | 'A COMMANDER' *(can be encoded as a sentence embedding)* |

# Contents

# Definition modeling with Flan-T5

- We generate definitions by prompting a (fine-tuned) `Flan-T5` language model [Chung et al., 2022].
- Fine-tuning is done in a straightforward seq2seq setup:
- *'Up until the middle of last century farmers were limited to cutting the hedges back with a hand slasher. What is slasher?'*
  - 'A TOOL FOR CUTTING VEGETATION WITH A LONG, SHARP BLADE'

# Definition modeling with Flan-T5

- We generate definitions by prompting a (fine-tuned) `Flan-T5` language model [Chung et al., 2022].
- Fine-tuning is done in a straightforward seq2seq setup:
- *'Up until the middle of last century farmers were limited to cutting the hedges back with a hand slasher. What is slasher?'*
  - 'A TOOL FOR CUTTING VEGETATION WITH A LONG, SHARP BLADE'

## Definition datasets for English (target word, definition, usage example)

| Dataset | Entries | Lemmas | Ratio | Usage length | Defin. length |
|---|---|---|---|---|---|
| **WordNet** [Ishiwatari et al., 2019] | 15,657 | 8,938 | 1.75 | $4.80^{\pm 3.43}$ | $6.64^{\pm 3.77}$ |
| **Oxford** [Gadetsky et al., 2018] | 122,318 | 36,767 | 3.33 | $16.73^{\pm 9.53}$ | $11.01^{\pm 6.96}$ |
| **CoDWoE** [Mickus et al., 2022] | 63,596 | 36,068 | 2.44 | $24.04^{\pm 21.05}$ | $11.78^{\pm 8.03}$ |

# Evaluation setup

- Target lemmas and usage examples from the definitions datasets

# Evaluation setup

- Target lemmas and usage examples from the definitions datasets
- conditionally generate definitions with `Flan-T5`
  - (no tweaking done, dumb greedy search with target word filtering to avoid circular definitions)

# Evaluation setup

- Target lemmas and usage examples from the definitions datasets
- conditionally generate definitions with `Flan-T5`
  - (no tweaking done, dumb greedy search with target word filtering to avoid circular definitions)
- compare generated definitions to the gold ones in the datasets, using reference-based NLG metrics.

# Evaluation setup

- Target lemmas and usage examples from the definitions datasets
- conditionally generate definitions with `Flan-T5`
  - (no tweaking done, dumb greedy search with target word filtering to avoid circular definitions)
- compare generated definitions to the gold ones in the datasets, using reference-based NLG metrics.

## Generalization tests

Following the GenBench generalisation taxonomy [Hupkes et al., 2022]:

1. Zero-shot: no fine-tuning, LLM as is
2. In-distribution: LLM fine-tuned and tested on the same dataset
3. Hard domain shift: LLM fine-tuned on dataset **A**, tested on dataset **B**
4. Soft domain shift: LLM fine-tuned on **all** datasets, tested on **one**.

# Simple and efficient approach

- The resulting definition generator performs on par with prior work
- ... but much simpler and more efficient.

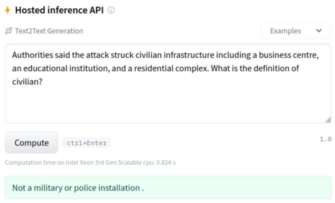
| | | *WordNet test set* | | | *Oxford test set* | | |
|---|---|---|---|---|---|---|---|
| Model | Generalization test | BLEU | ROUGE-L | BERT-F1 | BLEU | ROUGE-L | BERT-F1 |
| [Huang et al., 2021] | In-distribution | 32.72 | - | - | **26.52** | - | - |
| Flan-T5 XL | Zero-shot (task shift) | 2.70 | 12.72 | 86.72 | 2.88 | 16.20 | 86.52 |
| Flan-T5 XL | In-distribution | 11.49 | 28.96 | 88.90 | 16.61 | 36.27 | 89.40 |
| Flan-T5 XL | Hard domain shift | 29.55 | 48.17 | 91.39 | 8.37 | 25.06 | 87.56 |
| Flan-T5 XL | Soft domain shift | **32.81** | **52.21** | **92.16** | 18.69 | **38.72** | **89.75** |

# Simple and efficient approach

- The resulting definition generator performs on par with prior work
- ... but much simpler and more efficient.

| | | *WordNet test set* | | | *Oxford test set* | | |
|---|---|---|---|---|---|---|---|
| Model | Generalization test | BLEU | ROUGE-L | BERT-F1 | BLEU | ROUGE-L | BERT-F1 |
| [Huang et al., 2021] | In-distribution | 32.72 | - | - | **26.52** | - | - |
| Flan-T5 XL | Zero-shot (task shift) | 2.70 | 12.72 | 86.72 | 2.88 | 16.20 | 86.52 |
| Flan-T5 XL | In-distribution | 11.49 | 28.96 | 88.90 | 16.61 | 36.27 | 89.40 |
| Flan-T5 XL | Hard domain shift | 29.55 | 48.17 | 91.39 | 8.37 | 25.06 | 87.56 |
| Flan-T5 XL | Soft domain shift | **32.81** | **52.21** | **92.16** | 18.69 | **38.72** | **89.75** |

Try it yourself:
`https://huggingface.co/ltg/flan-t5-definition-en-large`
*(the Large model is 780M parameters, XL model is 3B)*



# Definitions in word-in-context similarity task

- Generated definitions allow quantitative comparisons between words in context:
  - *‘He went to the ball and polked himself into he good graces of Miss Juliet Trevor’*
  - *‘The big man threw the first two balls very hard anf fast’*
  - Semantic proximity: 1 out of 4.

# Definitions in word-in-context similarity task

- Generated definitions allow quantitative comparisons between words in context:
  - *‘He went to the ball and polked himself into he good graces of Miss Juliet Trevor’*
  - *‘The big man threw the first two balls very hard anf fast’*
  - Semantic proximity: 1 out of 4.
- One can compare definitions directly as strings:
  - Exact match
  - Levenstein distance
  - BLEU
  - METEOR
  - etc

# Definitions in word-in-context similarity task

- Generated definitions allow quantitative comparisons between words in context:
  - *'He went to the ball and polked himself into he good graces of Miss Juliet Trevor'*
  - *'The big man threw the first two balls very hard anf fast'*
  - Semantic proximity: 1 out of 4.
- One can compare definitions directly as strings:
  - Exact match
  - Levenstein distance
  - BLEU
  - METEOR
  - etc
- ...or compute cosine between definitions vectorized with SBERT [Reimers and Gurevych, 2019]

# Definitions in word-in-context similarity task

- Generated definitions allow quantitative comparisons between words in context:
  - *'He went to the ball and polked himself into he good graces of Miss Juliet Trevor'*
  - *'The big man threw the first two balls very hard anf fast'*
  - Semantic proximity: 1 out of 4.
- One can compare definitions directly as strings:
  - Exact match
  - Levenstein distance
  - BLEU
  - METEOR
  - etc
- ...or compute cosine between definitions vectorized with SBERT [Reimers and Gurevych, 2019]
- We tested it on diachronic word usage graphs (DWUGs) for English [Schlechtweg et al., 2021]

# Definitions in word-in-context similarity task

- Pairwise similarities between definitions correlate with human semantic similarity judgements better than token and sentence embeddings:

| Method | Cosine | SacreBLEU | METEOR |
|---|---|---|---|
| `RoBERTa-large` token embeddings | 0.141 | - | - |
| `SBERT` sentence embeddings | 0.114 | - | - |
| **Generated definitions** | | | |
| FLAN-T5 XL zero-shot | 0.188 | 0.041 | 0.083 |
| FLAN-T5 XXL zero-shot | 0.206 | 0.045 | 0.092 |
| FLAN-T5 Base fine-tuned | 0.221 | 0.078 | 0.077 |
| FLAN-T5 XL fine-tuned | **0.264** | **0.108** | **0.117** |

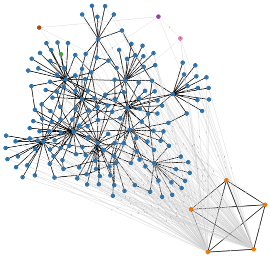
*Spearman correlations with human judgements*

# Contents

# Labeling word sense clusters with definitions

## Defining a collection of usages

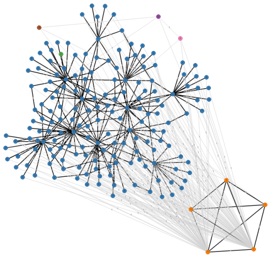
- Given: Several usage examples for a target word with data-driven usage clusters (senses)
- ... e.g., from the same DWUGs [Schlechtweg et al., 2021].
- We generate definitions for each usage, and find the most prototypical definitions.
- Human-readable sense labels instead of anonymous cluster ids!



# Labeling word sense clusters with definitions

## Defining a collection of usages

- Given: Several usage examples for a target word with data-driven usage clusters (senses)
- ... e.g., from the same DWUGs [Schlechtweg et al., 2021].
- We generate definitions for each usage, and find the most prototypical definitions.
- Human-readable sense labels instead of anonymous cluster ids!

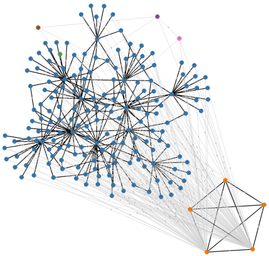


- One can also use the definition of the most prototypical usage
  - ... based on token embeddings

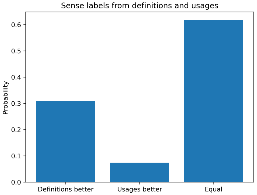
# Labeling word sense clusters with definitions

## Defining a collection of usages

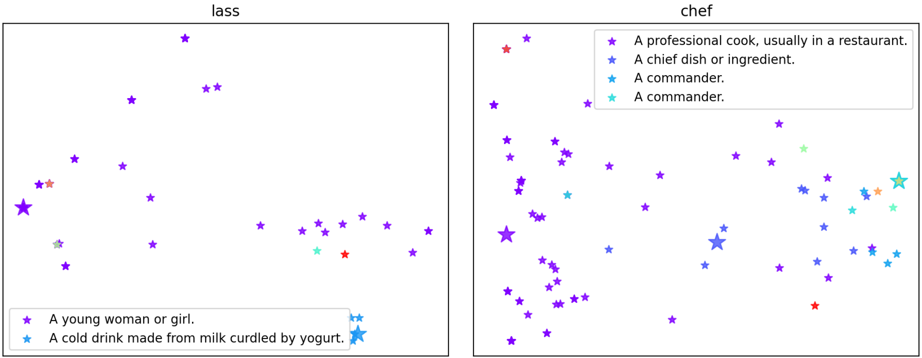
- Given: Several usage examples for a target word with data-driven usage clusters (senses)
- ... e.g., from the same DWUGs [Schlechtweg et al., 2021].
- We generate definitions for each usage, and find the most prototypical definitions.
- Human-readable sense labels instead of anonymous cluster ids!



- One can also use the definition of the most prototypical usage
  - ... based on token embeddings
- But human evaluation shows most prototypical definitions are consistently better.



# Labeling word sense clusters with definitions




PCA projections of definition embeddings for target words from English DWUG *(colors are data-driven sense clusters, large stars are prototypical definitions)*.

# Contents

# Explainable semantic change detection

## How related are our senses?

- Given a DWUG, we measure cosine similarities between labels of clusters/senses in all time periods.
- Most labels are very dissimilar, but some are unusually close to each other:
- for each target word, we simply find outlier pairs with $z > 1$.

# Explainable semantic change detection

## How related are our senses?

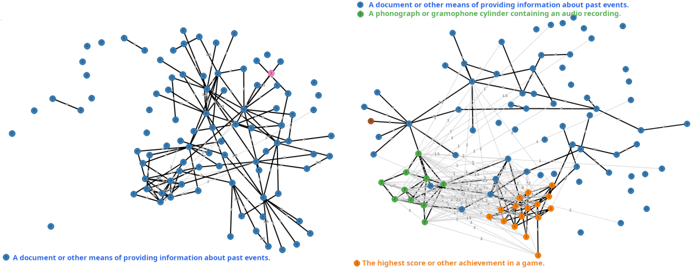
- Given a DWUG, we measure cosine similarities between labels of clusters/senses in all time periods.
- Most labels are very dissimilar, but some are unusually close to each other:
- for each target word, we simply find outlier pairs with $z > 1$.

This gives us an interpretable sense dynamics map:

- senses transitioning one into another
- splitting from another sense
- two senses merging into one
- etc.

# Explainable semantic change detection

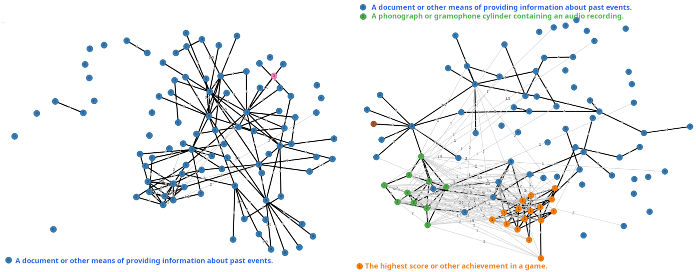
- Diachronic map: ‘a novel sense 2 of ‘record’ in time period 2 (‘A PHONOGRAPH OR GRAMOPHONE CYLINDER...’) is probably an offshoot of a stable sense 0 present in both time periods (‘A DOCUMENT OR OTHER MEANS OF PROVIDING INFORMATION...’)’ (narrowing)




*Left: time period 1 (1810-1860); right: time period 2 (1960-2010).*

# Explainable semantic change detection

- Diachronic map: ‘a novel sense 2 of ‘record’ in time period 2 (‘A PHONOGRAPH OR GRAMOPHONE CYLINDER...’) is probably an offshoot of a stable sense 0 present in both time periods (‘A DOCUMENT OR OTHER MEANS OF PROVIDING INFORMATION...’)’ (narrowing)




*Left: time period 1 (1810-1860); right: time period 2 (1960-2010).*

- Sense labels help to generate explanations of semantic change;
- ... actually useful for historical linguists, lexicographers, or social scientists

# Explainable semantic change detection

## Fixing DWUGs

- trace incorrect or inconsistent DWUG clustering
- Two sense clusters have the same label? Likely, they are one cluster/sense.

# Explainable semantic change detection

## Fixing DWUGs

- trace incorrect or inconsistent DWUG clustering
- Two sense clusters have the same label? Likely, they are one cluster/sense.

## 'Ball' example

- Sense similarities are non-transitive:
  - Ball 0: 'A SPHERE OR OTHER OBJECT USED AS THE OBJECT OF A HIT'
  - Ball 2: 'A ROUND SOLID PROJECTILE, SUCH AS IS USED IN SHOOTING'
  - Ball 3: 'A BULLET'
- $c_0$ to $c_2$: 0.70
- $c_2$ to $c_3$: 0.53
- $c_0$ to $c_3$: 0.50 (below the outlier threshold)

Inconsistent clustering, but also interesting insights about meaning trajectory of '*ball*'.

# Contents

# Definitions as representations

- Semantic change modelling is only one use case.

# Definitions as representations

- Semantic change modelling is only one use case.
- Our 'definitions as lexical representations' paradigm is promising for many NLP tasks.
- ...actually, [Bevilacqua et al., 2020] already employed definitions for WSD.

# Definitions as representations

- Semantic change modelling is only one use case.
- Our ‘definitions as lexical representations’ paradigm is promising for many NLP tasks.
- ...actually, [Bevilacqua et al., 2020] already employed definitions for WSD.
- Benefits:
  - human-readable representations
  - more abstract and robust to noise
  - outperforms ‘standard’ embeddings in word-in-context similarity judgements
  - for humanities, it’s easier to operate in the space of the definitions.

More in the paper:

`https://arxiv.org/abs/2305.11993`

# Full results

| | | WordNet test set | | | Oxford test set | | |
|---|---|---|---|---|---|---|---|
| Model | Generalization test | BLEU | ROUGE-L | BERT-F1 | BLEU | ROUGE-L | BERT-F1 |
| [Huang et al., 2021] | *Unknown* | 32.72 | - | - | **26.52** | - | - |
| T5 base | Zero-shot (task shift) | 2.01 | 8.24 | 82.98 | 1.72 | 7.48 | 78.79 |
| T5 base | Soft domain shift | 9.21 | 25.71 | 86.44 | 7.28 | 24.13 | 86.03 |
| Flan-T5 base | Zero-shot (task shift) | 4.08 | 15.32 | 87.00 | 3.71 | 17.25 | 86.44 |
| Flan-T5 base | In-distribution | 8.80 | 23.19 | 87.49 | 6.15 | 20.84 | 86.48 |
| Flan-T5 base | Hard domain shift | 6.89 | 20.53 | 87.16 | 4.32 | 17.00 | 85.88 |
| Flan-T5 base | Soft domain shift | 10.38 | 27.17 | 88.22 | 7.18 | 23.04 | 86.90 |
| Flan-T5 large | Soft domain shift | 14.37 | 33.74 | 88.21 | 10.90 | 30.05 | 87.44 |
| T5 XL | Zero-shot (task shift) | 2.05 | 8.28 | 81.90 | 2.28 | 9.73 | 80.37 |
| T5 XL | Soft domain shift | **34.14** | **53.55** | 91.40 | 18.82 | 38.26 | 88.81 |
| Flan-T5 XL | Zero-shot (task shift) | 2.70 | 12.72 | 86.72 | 2.88 | 16.20 | 86.52 |
| Flan-T5 XL | In-distribution | 11.49 | 28.96 | 88.90 | 16.61 | 36.27 | 89.40 |
| Flan-T5 XL | Hard domain shift | 29.55 | 48.17 | 91.39 | 8.37 | 25.06 | 87.56 |
| Flan-T5 XL | Soft domain shift | **32.81** | **52.21** | **92.16** | 18.69 | **38.72** | **89.75** |

# References I


Bevilacqua, M., Maru, M., and Navigli, R. (2020).
Generationary or "how we went beyond word sense inventories and learned to gloss".
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022).
Scaling instruction-finetuned language models.
*arXiv preprint arXiv:2210.11416*.

# References II


Gadetsky, A., Yakubovskiy, I., and Vetrov, D. (2018).
Conditional generators of words definitions.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Huang, H., Kajiwara, T., and Arase, Y. (2021).
Definition modelling for appropriate specificity.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# References III


Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., et al. (2022).
State-of-the-art generalisation research in NLP: A taxonomy and review.
*arXiv preprint arXiv:2210.03050*.

Ishiwatari, S., Hayashi, H., Yoshinaga, N., Neubig, G., Sato, S., Toyoda, M., and Kitsuregawa, M. (2019).
Learning to describe unknown phrases with local and global contexts.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

# References IV


Mickus, T., Van Deemter, K., Constant, M., and Paperno, D. (2022).
Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings.
In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019).
Sentence-BERT: Sentence embeddings using Siamese BERT-networks.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

# References V

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021).
DWUG: A large resource of diachronic word usage graphs in four languages.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.