Building Semantic Frame Resources Using Large Language Models

Ryohei Sasano (Nagoya University)

The slides can be downloaded via this QR code

NAGOYA UNIVERSITY

Does LLM understand semantic frames?

FINISH_COMPETITION

A Competition comes to an end, with a Competitor tying, winning, or losing against an Opponent …

He **lost** the gold medal by just .02 points.

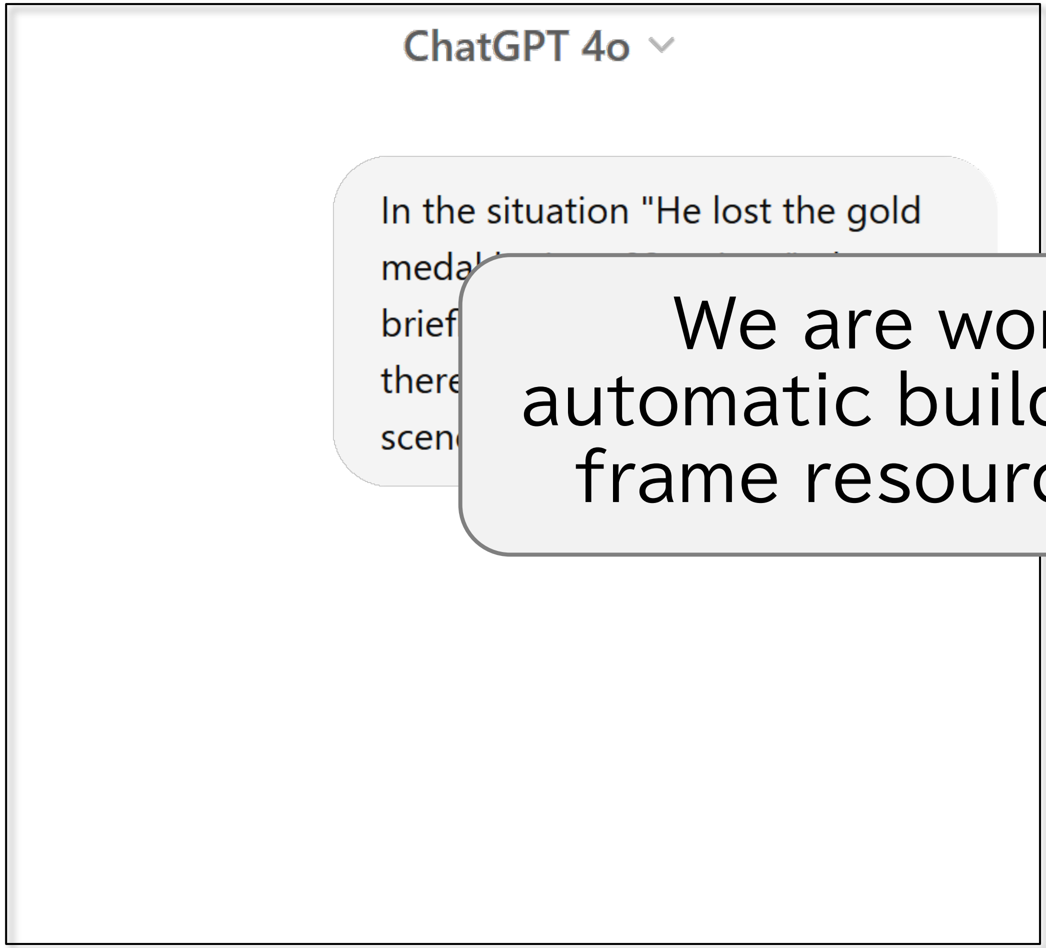
He **lost** his gold medal at the restaurant.

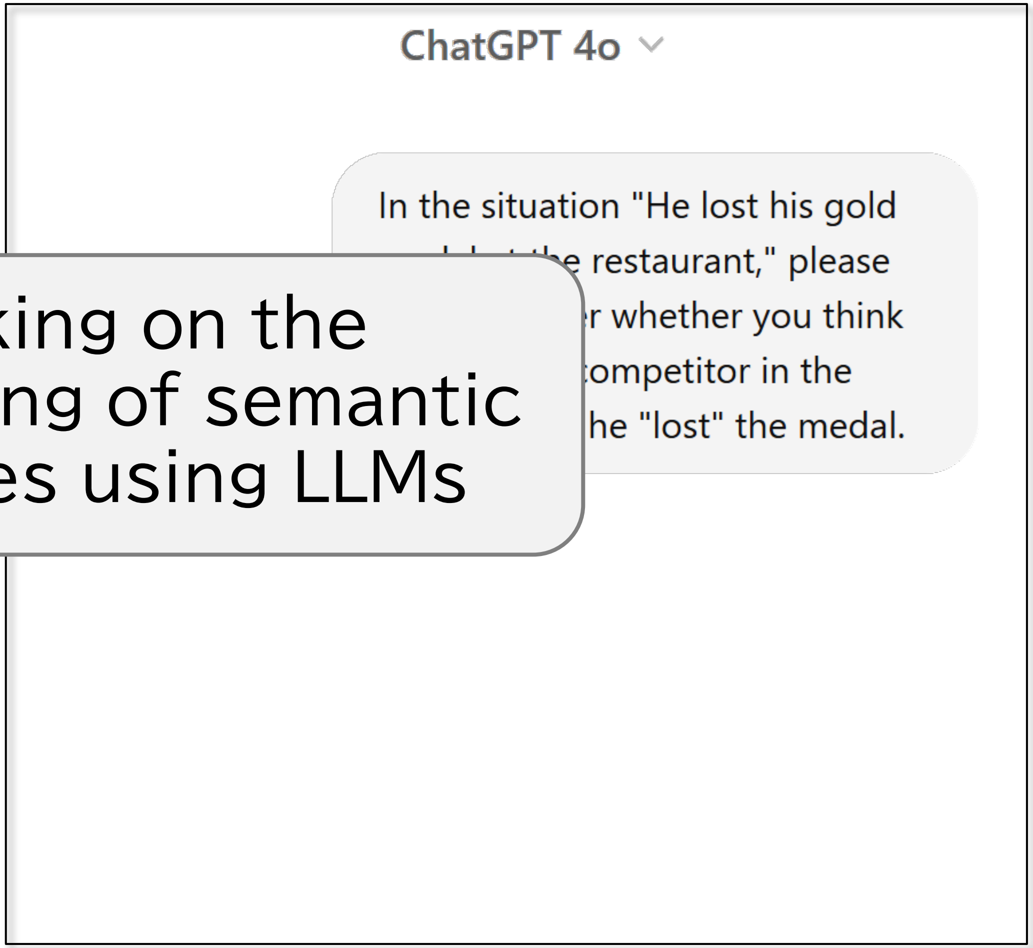
LOSING

This frame describes a situation in which an Owner loses his or her Possession …

The slides can be downloaded via this QR code

LLMs are likely to have a good understanding of semantic frames





We are working on the automatic building of semantic frame resources using LLMs

Building Frame Resources using LLMs

- What are LLMs?
  - Language models with many parameters, which are trained with self-supervised learning on a vast amount of text
  - In this presentation, LLMs include not only recent causal LMs such as GPT and Llama, but also masked LMs such as BERT

- Why do we build frames automatically?
  1. To make it easier to develop semantic frame resources tailored to specific languages, domains, and other objectives
  2. To support for the manual development of frame resources
  3. To analyze how close to human the LLM understands language

Contents

Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering

Kosuke Yamada, Ryohei Sasano, Koichi Takeda [In Proc. of ACL-IJCNLP 2021]

Semantic Frame Induction

- Clustering frame-evoking words according to the semantic frames they evoke
- In the following example, the goal is to group {1,2}, {3}, and {4} together

1. He acquires marketing data .
2. She gets a doll for Christmas .

Getting

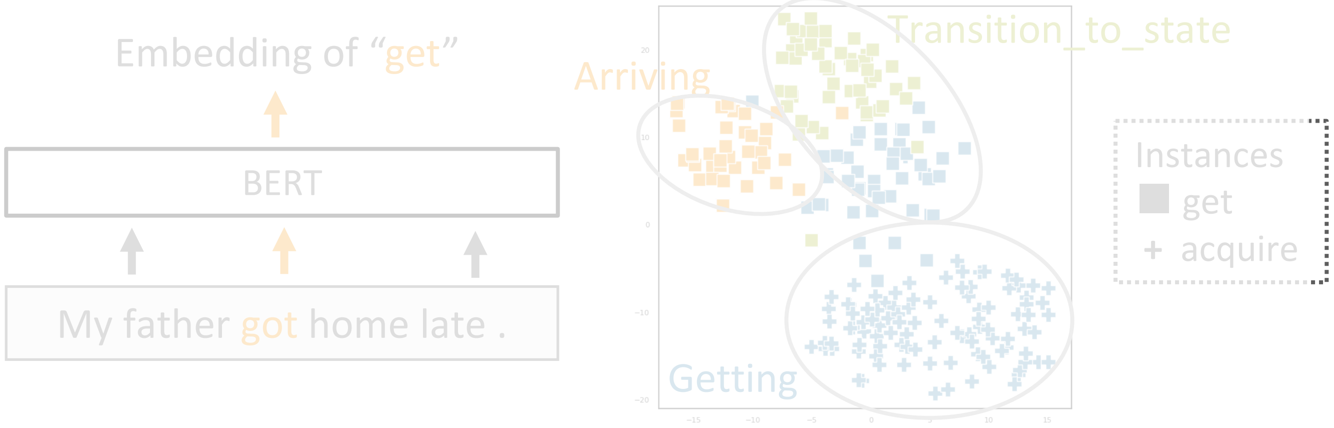
3. My father got home late .

Arriving

4. The situation got worse .

Transition_to_state

- One of the tasks in SemEval-2019 Task 2: Unsupervised Lexical Frame Induction
  - Following the shared task settings, we only consider verbs as frame-evoking words
  - Top three methods leverage contextualized word embeddings such as BERT for clustering



Related Work on Frame Induction

- Methods not using contextualized word embeddings
  - [Kawahara+’14] first extracted predicate-arguments structures and then use the Chinese Restaurant Process to group verbs
  - [Ustalov+’18] perform graph-based clustering using concatenated embeddings of static embeddings of verb, subject, and object

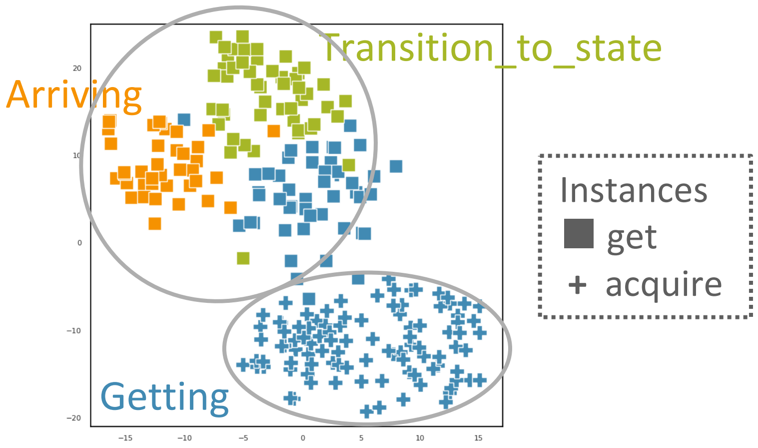
- Methods using contextualized word embeddings
  - [Anwar+’19] perform group average clustering using ELMo embeddings
  - [Ribeiro+’19] perform graph-based clustering based on Chinese whispers by using ELMo embeddings
  - [Arefyev+’19] first perform group average clustering using BERT embeddings, and then split each cluster into two

Two problems and solutions

When using BERT embeddings for frame induction, there are two problems

Problem 1: Examples of the same verb tends to be distributed nearby

Solution 1: Using masked word embeddings

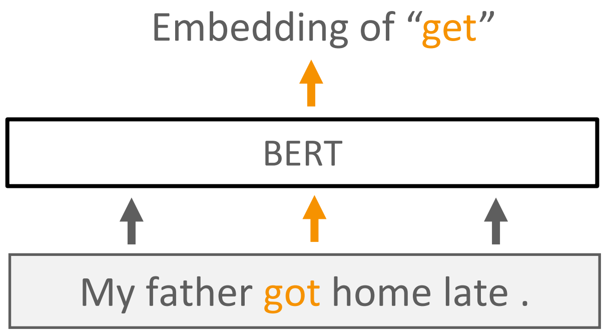


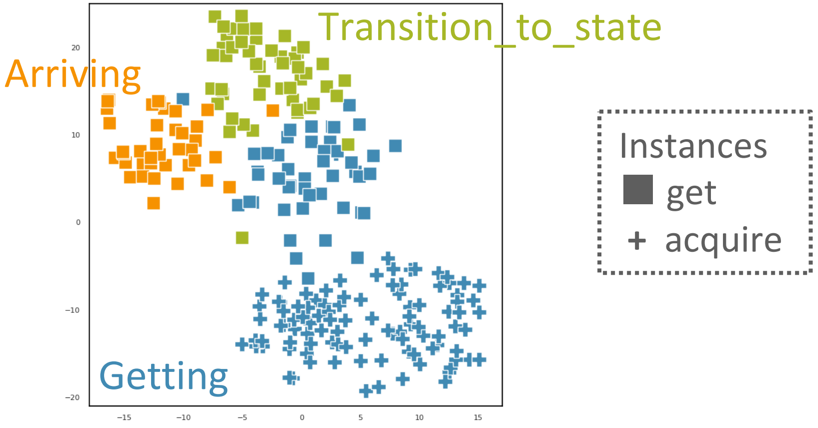
Problem 2: If instances of all verbs are clustered simultaneously, the instances of the same verb tend to be split into too many different clusters

Solution 2: Two-step clustering: first, clustering is conducted within a verb, followed by clustering across verbs

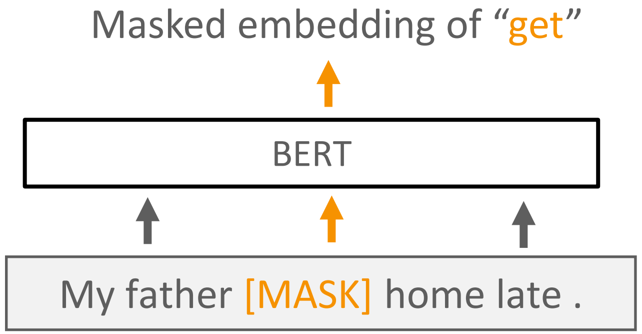
Solution 1: Using the mask word embedding to suppress the surface information of the verb

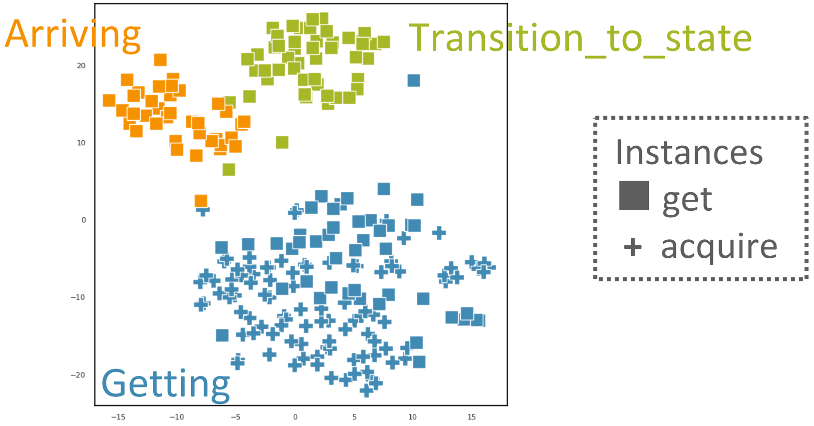
- Normal word embeddings of BERT





- Masked word embeddings of BERT

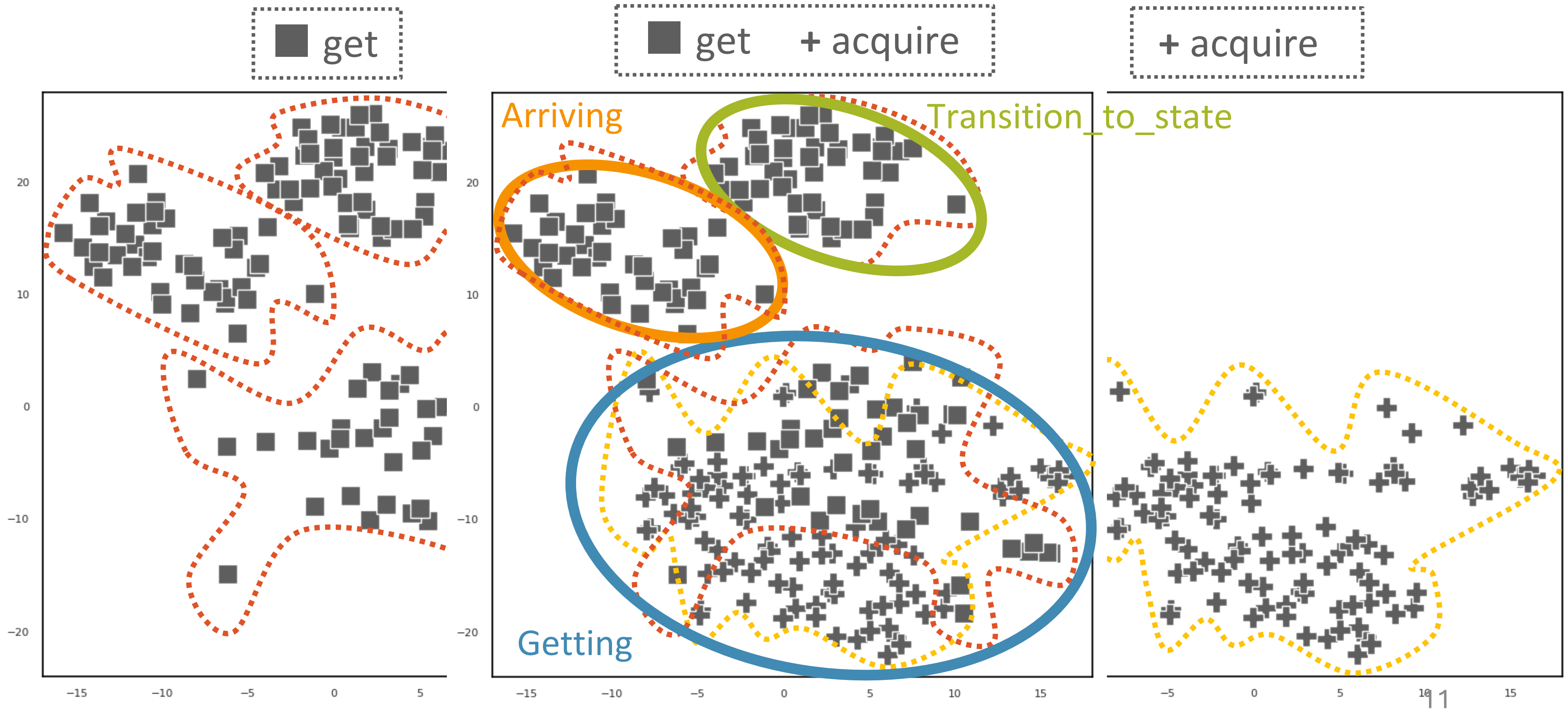




- We use $v_{w+m} = \alpha \cdot v_{mask} + (1 - \alpha) \cdot v_{word}$

Solution 2: Two-step clustering

- 1st step: Clustering Instances of the Same Verb
- 2nd step: Clustering across Verbs



Evaluation of Frame Induction

- Evaluation
  - We use the manually frame-annotated data as reference
  - We use the same metrics as SemEval-2019 (Task 2):
    - Purity (PU), inverse-Purity (IPU), and their harmonic mean (PIF)
    - B-Cubed Precision (BcP), Recall (BcR), and their harmonic mean (BcF)

- Data for evaluation
  - The SemEval-2019 (Task 2) dataset is not publicly available
  - We extracted verbal LUs with at least 20 example sentences from FrameNet 1.7

	#Verbs	#LUs	#Frames	#Examples
Dev.	255	300	169	12,718
Test	1,017	1,188	393	47,499
All	1,272	1,488	434	60,217

Statistics of the dataset from FrameNet

Experimental Results

Model	Clustering		α	PU / IPU / PIF	BCP / BCR / BCF
1-cluster-per-head	1cpv		–	88.9 / 39.7 / 54.9	86.6 / 33.9 / 48.7
Arefyev et al. (2019)	GA (Cosine)		–	69.9 / 55.1 / 61.6	62.8 / 44.0 / 51.7
Anwar et al. (2019)	GA (Manhattan)		–	71.5 / 52.0 / 60.2	65.1 / 41.0 / 50.3
Ribeiro et al. (2019)	Chinese Whispers		–	50.9 / 66.3 / 57.5	39.4 / 56.7 / 46.5
One-step clustering	Ward		0.0	64.3 / 49.5 / 56.0	55.2 / 38.9 / 45.6
	GA		0.0	38.7 / 64.9 / 48.5	26.1 / 52.5 / 34.9
	first-step	**second-step**			
Two-step clustering	GA	Ward	0.9	49.3 / 72.9 / 58.8	37.3 / 64.6 / 47.3
	GA	GA	0.6	63.0 / **76.3** / 69.0	52.8 / **68.0** / 59.4
	X-means	Ward	0.8	54.0 / 72.2 / 61.8	42.6 / 63.6 / 51.1
	X-means	GA	0.7	**71.9** / 74.1 / **73.0**	**63.2** / 65.5 / **64.4**

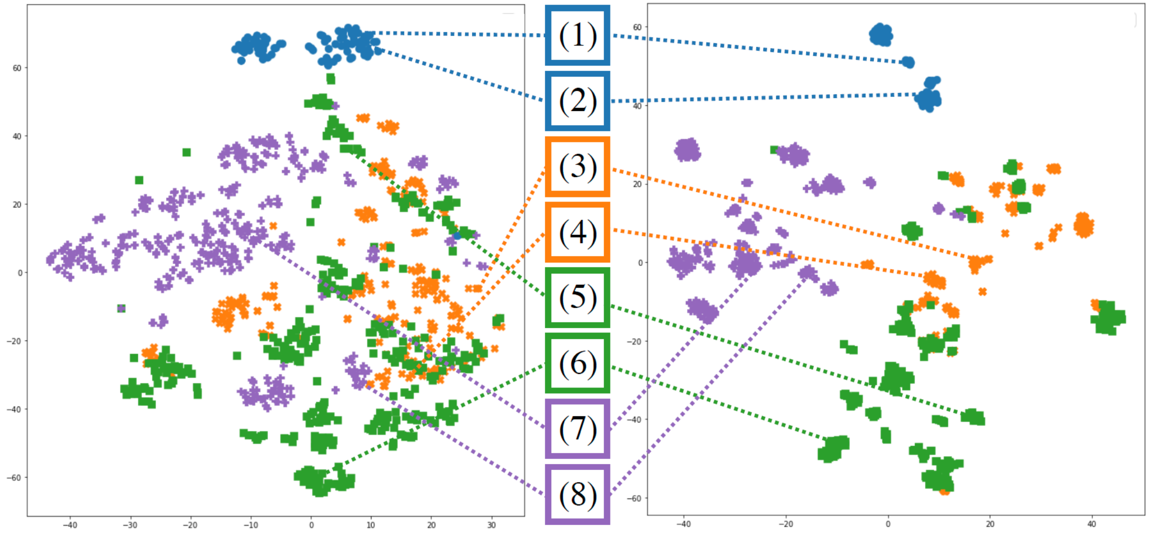
cf. $v_{w+m} = \alpha \cdot v_{mask} + (1 - \alpha) \cdot v_{word}$

Semantic Frame Induction with Deep Metric Learning

Kosuke Yamada, Ryohei Sasano, Koichi Takeda [In Proc. of EACL 2023]

Fine-tuning BERT using deep metric learning to align with human intuition for frames

- We confirmed that BERT contains knowledge on semantic frames
- However, BERT embeddings reflect various aspects of words, and those related to frames are only a part of it
- We fine-tune BERT using deep metric learning (DML) to optimize it for frame knowledge and use it for frame induction
- Note that this method assumes manually annotated information is available for some frames



(a) Vanilla BERT (b) Fine-tuned BERT w/ ArcFace

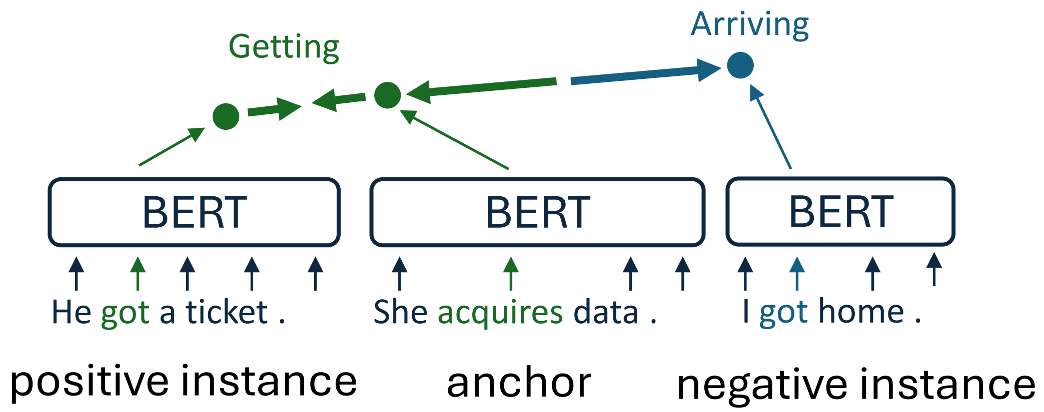
(1) Chapter 8 treats the educational advantages.
(2) Each database will cover a specific topic.
(3) She covered her mouth with her hand.
(4) I filled a notebook with my name.
(5) You can embed graphs in your worksheet.
(6) He parked the car at the hotel.
(7) Volunteers removed grass from the marsh.
(8) They'd drained the last drop from the teapot.

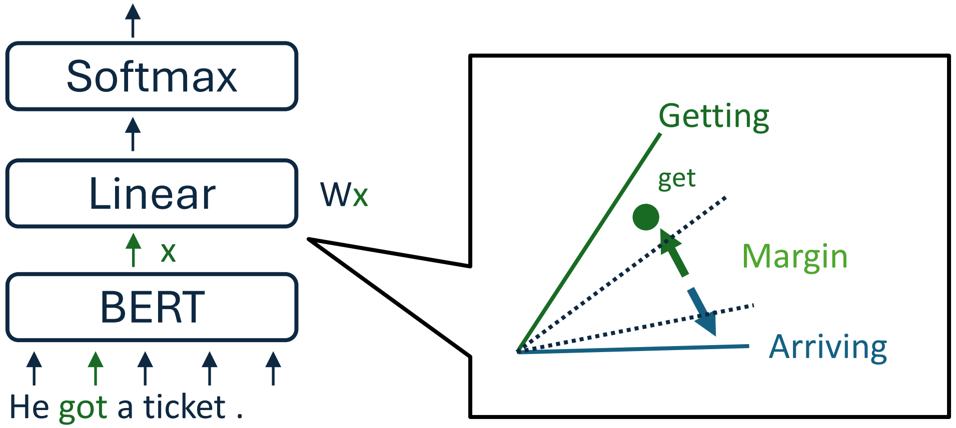
● TOPIC (1)–(2)
✖ FILLING (3)–(4)
■ PLACING (5)–(6)
✚ REMOVING (7)–(8)

Fine-tuning BERT via DML

- We fine-tune BERT so that instances of verbs that evoke the same frame are closer together, and others are further apart

- We adopt several representative loss functions
  - Triplet loss:
    - Fine-tuning so that the distance from the anchor to the negative instance is more than a certain margin away from the distance to the positive instance
  - ArcFace & AdaCos:
    - Classification-based losses, which has recently become the standard for face recognition





Experiments

- Dataset
  - The instances extracted from FrameNet 1.7 were split into three sets so that sentences with the same verb were in the same set
  - We performed three-fold cross validation with the three sets as the training, development, and test sets

- Induction Model
  - We used [Yamada+'21] as a baseline model, which is not perform finetuning (vanilla)

	#Verbs	#LUs	#Frames	#Instances
Set 1	831	1,277	429	28,314
Set 2	831	1,261	415	26,179
Set 3	830	1,280	459	28,117
All	2,492	3,818	642	82,610

Experimental Results

cf. $v_{w+m} = \alpha \cdot v_{mask} + (1 - \alpha) \cdot v_{word}$

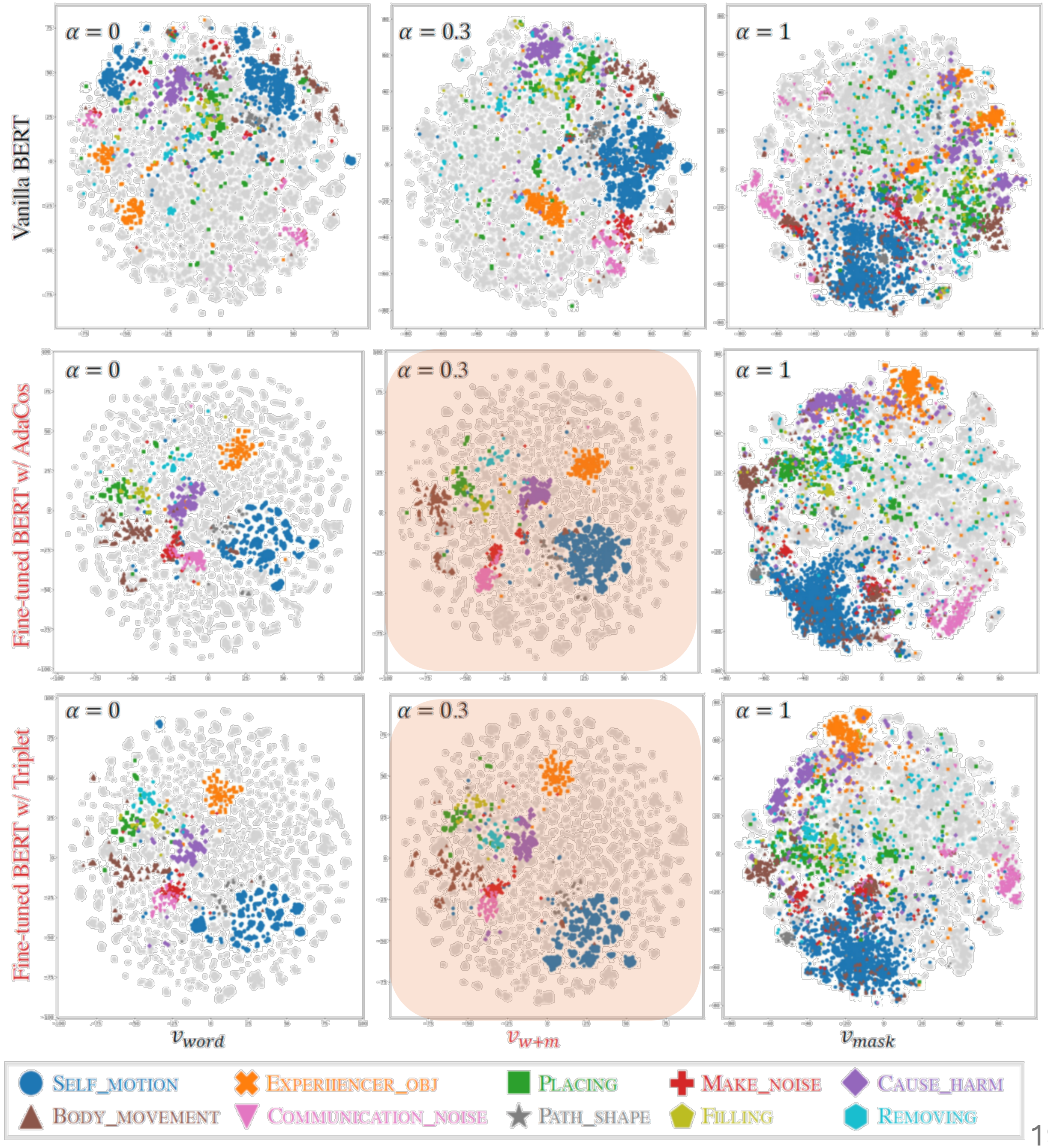
Clustering	Model	α	PU / IPU / PIF	BCP / BCR / BCF
One-step clustering	Vanilla	0.00	53.0 / 57.0 / 54.9	40.8 / 44.6 / 42.6
	Triplet	0.23	70.0 / 77.0 / 73.3	**60.3** / 68.1 / **63.9**
	ArcFace	0.37	**70.3** / 76.2 / 73.1	59.7 / 67.4 / 63.3
	AdaCos	0.30	69.0 / **78.7** / **73.5**	57.5 / **69.5** / 62.9
Two-step clustering	Vanilla	0.67	60.6 / 74.9 / 66.9	49.7 / 65.8 / 56.5
	Triplet	0.50	73.4 / **76.7** / 74.8	64.6 / **68.0** / 66.0
	ArcFace	0.47	70.5 / 76.5 / 73.3	60.8 / 67.7 / 63.8
	AdaCos	0.50	80.8 / 71.3 / **75.6**	73.2 / 60.9 / **66.2**

- Performance is greatly improved by fine-tuning via DML
  - Compared to Vanilla, accuracy has improved by almost points
- Differences between one-step and two-step clustering has become small

Visualization of Embeddings

- We make 2D t-SNE projections of v_{word}, v_{w+m}, v_{mask} for the Vanilla, AdaCos, and Triplet models
- All verbs are mapped in two dimensions, and the top 10 frames by frequency are colored
- After fine-tuning, examples belonging to the same frame are clustered together

- The approach using deep metric learning is also very effective in argument clustering [Yamada+' 23b]



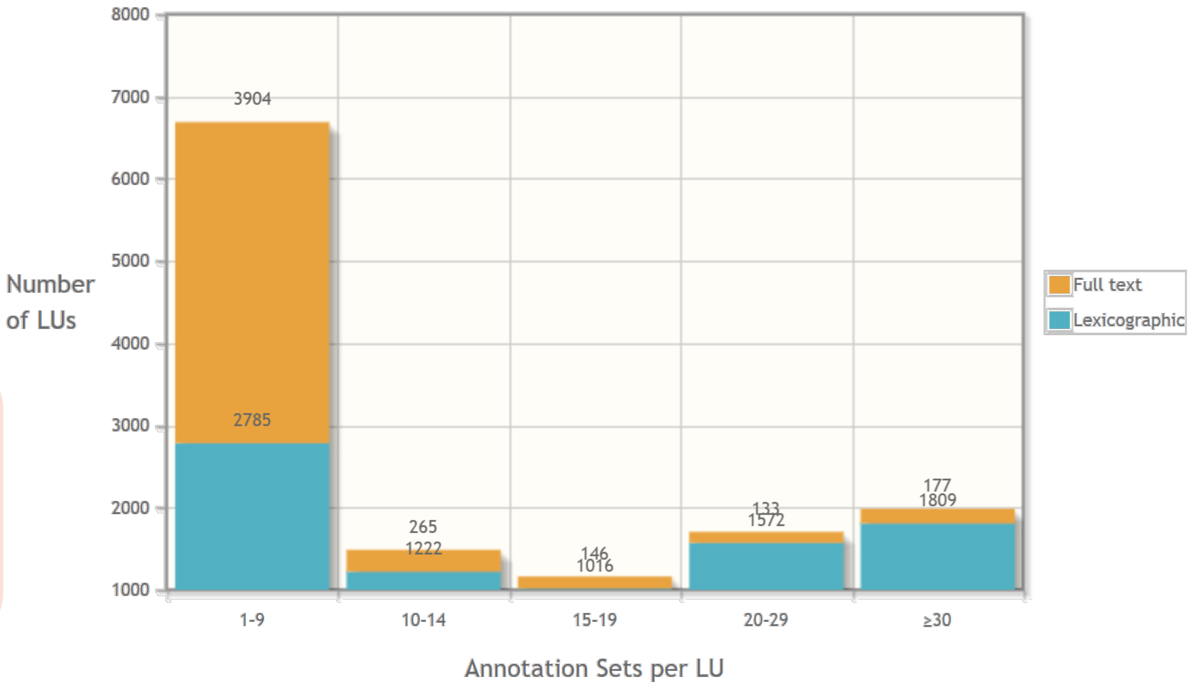
Semantic Frame Induction from Real Corpora

Shogo Tsujimoto, Kosuke Yamada, Ryohei Sasano, [In progress (IPSJ-SIGNL)]

Two types of annotation sets in FrameNet

- Lexicographic annotation set
  - Sentences are chosen because they contain a predetermined target LU
  - Annotation is done relative to only one lexical unit per sentence
- Full text annotation set
  - All sentences in a given text are the target of annotation
  - All lexical units are treated as targets and their dependents are annotated
- Research Question
  - What happens when frame induction is done using a more realistic corpus, the Colossal Clean Crawled Corpus (C4)

Statistics
174532 (86%): Lexicographic
28446 (14%): Full Text



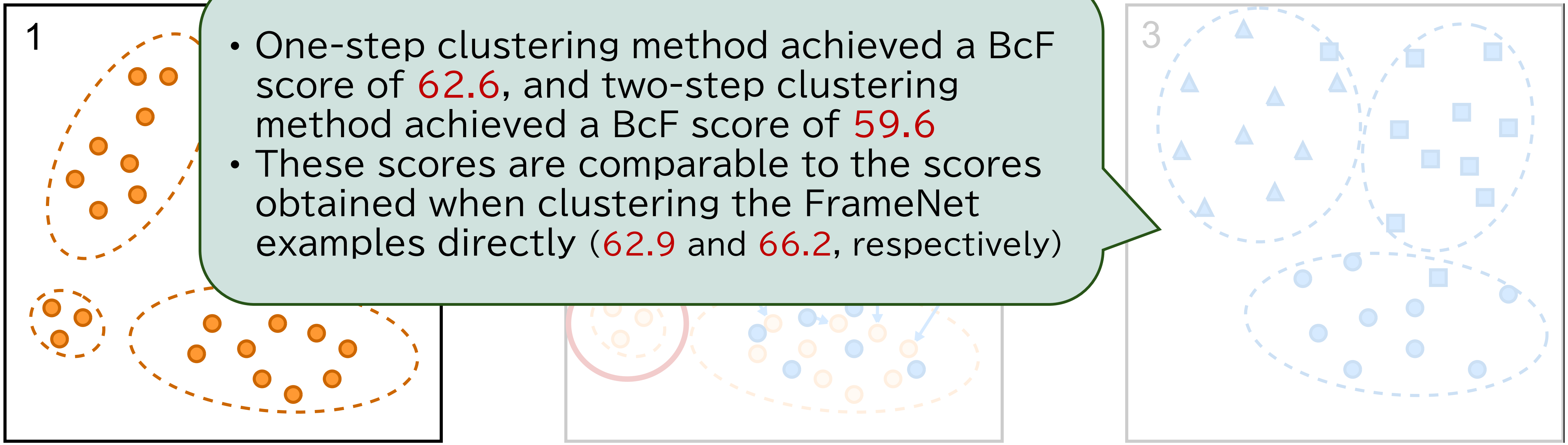
https://framenet.icsi.berkeley.edu/current_status (2024.05.19)

How does the distribution of examples in FrameNet differ from that of real-world corpora?

1. Recent texts are rarely included in the FrameNet examples
    - At least 89.2% of the examples were annotated before 2008
    - Examples of relatively new word meanings may not be included
2. The distribution of meanings for each verb differs from the distribution of meanings in actual corpora
    - 86% of the examples are included in the lexicographic annotation set
3. The distribution of annotated verbs differs from the actual distribution of verbs
    - In this study, to enable evaluation using FrameNet examples, the distribution of verb occurrences will be made the same (details on the next slide)

Flow of Experiment and Evaluation

1. Frame induction is performed using examples extracted from C4 with a constraint that the occurrence rate of verbs is aligned with FrameNet
2. Each FrameNet example used for evaluation is assigned to a cluster that contains the most similar examples
3. The assignments are regarded as the clustering results and evaluated



- One-step clustering method achieved a BcF score of 62.6, and two-step clustering method achieved a BcF score of 59.6
- These scores are comparable to the scores obtained when clustering the FrameNet examples directly (62.9 and 66.2, respectively)

● : example from C4

● ▲ ■ : examples in FrameNet, the shape represents the frame they evoke

Example of a cluster to which no FrameNet example was assigned

- Cluster 1

... should not **rush** a patient ...
Do not **rush** yourself!
... you do not **rush** this process
... being **hastened** ... by the ...

- In FrameNet, **rush** is the LU of the Fluidic_motion and the Self_motion frame, **hasten** is the LU of the Self_motion frame
- This cluster suggests the existence of frames not covered by FrameNet

- Cluster 2

... **stream** the video ...
... be **streamed** on 5G.
... can use it to **stream** music ...
... **stream** media and play games

- In FrameNet, **stream** is the LU of the Fluidic_motion and Mass_motion frame
- This cluster corresponds to the meaning that has become common in recent years
- This suggests the possibility of automatic acquisition of frames corresponding to new meanings

Definition Generation for Automatically Induced Semantic Frame

Yi Han, Ryohei Sasano, Koichi Takeda [In Findings of ACL 2024]

Frame Definition in FrameNet

- In FrameNet, a frame definition is a textual description of what the frame represents
- While the frame induction task provides clusters of frames, it lacks interpretability because definitions of these clusters are not provided
- To make frame resources intuitive and understandable to humans, we attempt to make frame definition

Cutting

Definition:

An Agent cuts a Item into Pieces using an Instrument (which may or may not be expressed).

At the ceremony, the CEO CUT the red ribbon hanging across the main entrance into a glorious confetti.

FEs:

Core:

Agent [Agt] **Semantic Type:** Sentient	The Agent is the person cutting the Item into Pieces.
Item [Item]	The item which is being cut into Pieces. People back then had to CHOP firewood all day long.
Pieces [Pie]	The Pieces are the parts of the original Item which are the result of the slicing.

Non-Core:

Instrument [Ins] **Semantic Type:** Physical_entity	The Instrument with which the Item is being cut into Pieces.
Manner [Manr] **Semantic Type:** Manner	Manner in which the Item is being cut into Pieces.
Means [Mns] **Semantic Type:** State_of_affairs	An act of the Agent that accomplishes the slicing. I SLICED the cucumber in 1/8th inch slices by marking intervals with a ruler.

Frame Definition Generation

- Input:
  - A set of frame-evoking words
  - Their exemplars
- Output:
  - A definition that accurately captures the essence of the frame they evoke

Frame: CUTTING

- **Frame definition**: Output

  An AGENT cuts an ITEM into PIECES using an instrument.
- **Frame elements (core)**:

  AGENT, ITEM, PIECES
- **Frame evoking words**: Input

  slice, cut, chop, dice, fillet, mince, · · ·
- **Exemplars**:
  - ◇ I carefully sliced the tomatoes for the salad.
  - ◇ She cut into the melon with a knife.
  - ◇ Chop the onions finely.

Leveraging In-context Learning

Task instruction

In the given exemplars, the words in each word list are used in similar context, suggesting a common underlying frame.Please provide me frame elements and a one-sentence brief definition for the frame that the words in word list n+1 evoke.

Demonstrations

× n

Query

Word list 1:
ransack, rob, mug
Exemplars:
'Ransack' in exemplar: ……
'rob' in exemplar: ……
'mug' in exemplar: ……

Output

Frame elements:
PERPETRATOR, VICTIM, SOURCE
Definition:
These are words describing situations in which a PERPETRATOR wrongs a VICTIM by taking something (Goods) from them.

New query (Input)

Word list n+1:
slice cut chop dice fillet mince pare carve cube
Exemplars:
'slice' in exemplar: I carefully sliced the tomatoes for the salad.
'cut ' in exemplar: She cut into the melon with a knife.
'chop' in exemplar: Chop the onions finely.

LLM

Output

Frame Elements :
AGENT, ITEM, PIECES
Definition:
An AGENT cuts an ITEM into PIECES using an instrument.

We used
Llama2-70b-chat

Def-Eval: Evaluating definitions with LLMs

Task instruction

You will evaluate a generated definition of a semantic frame. Provided **with the ground truth reference definition** of this frame, your task is to assess the definitions based on their ability to conclude the semantic frame. Please give me the number 1 to 5 directly following the criteria below.

Criteria

5: The two definitions are completely equivalent, as they mean the same thing.
4: The two definitions are mostly equivalent, but some unimportant details differ.
3: The two definitions are roughly equivalent, but some important information differs/missing.
2: The two definitions are not equivalent, but share some details.
1: The two definitions are completely dissimilar'

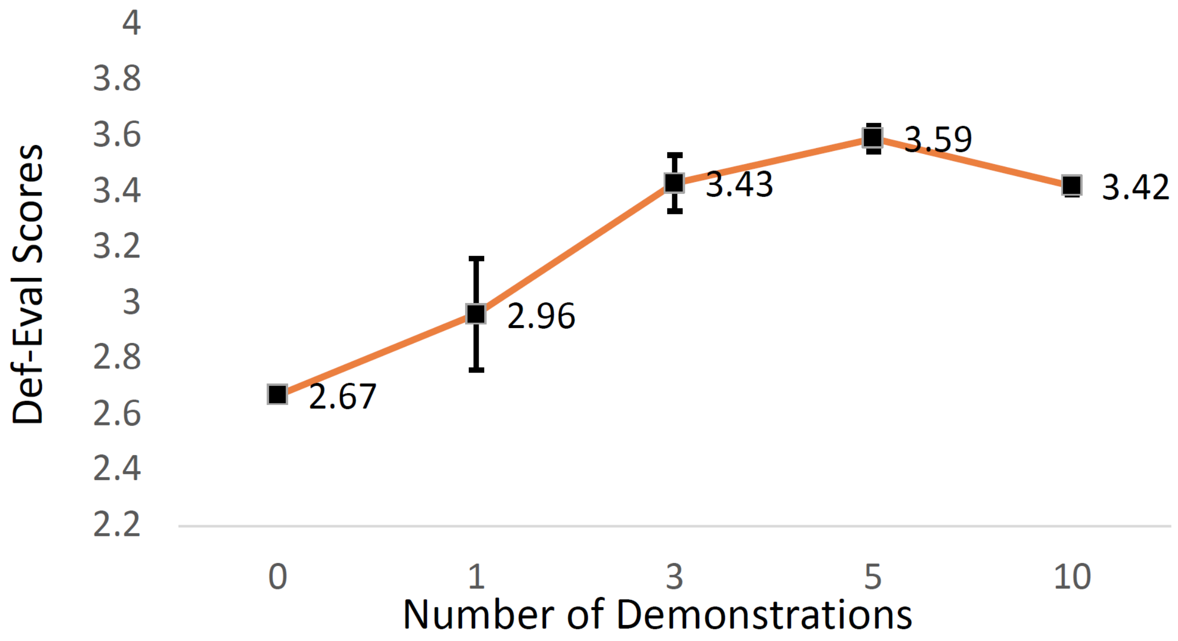
Reference definition and generated definition

Reference definition: An AGENT cuts an ITEM into PIECES using an instrument.
Generated definition: An AGENT divides SOMETHING with sharp tools.

GPT-4 → 3

The correlation with the manual evaluation is 0.48

Def-Eval Scores across Different Numbers of Demonstrations



- The number of demonstrations is suggested to be three or more
- The score is around 3.5, which means between mostly equivalent and roughly equivalent

5: The two definitions are completely equivalent, as they mean the same thing.
4: The two definitions are mostly equivalent, but some unimportant details differ.
3: The two definitions are roughly equivalent, but some important information differs/missing.
2: The two definitions are not equivalent, but share some details.
1: The two definitions are completely dissimilar'

Examples of Reference and Generated Definition and its Score

Frame	Definition (R=Reference, G=Generated)	Def-Eval
Judgment_ communication	**R:** A Communicator communicates a judgment of an Evaluee to an Addressee. **G:** An Agent expresses a strong evaluation or opinion about a Person, Action, or Thing, often with the intention of influencing others' opinions or behaviors.	3
Agriculture	**R:** In this frame, an Agriculturist cultivates Food. **G:** The Agent cultivates Crops on the Land using Tools, with the goal of producing food or other valuable resources.	4
Retaining	**R:** In this frame, an Agent continues to possess a Theme for a Duration of time. **G:** An Agent keeps or retains possession of an Object, maintaining its current State without any change or loss.	3

Summary & Future directions

- Summary
  - High-quality frame induction is possible by fine-tuning LLMs, such as BERT and Llama, if annotated resources are available for some frames
  - When frame induction is performed using the latest web corpus, frames not included in existing frame resources can be induced
  - LLM-based methods show potential for generating frame definition
- Future directions
  - Higher-quality frame induction is possible by using causal LMs with a larger number of parameters, such as Llama [Yano+, in progress]
  - Automatic recognition of inter-frame relationships using LLM
  - Leveraging large vision language model (LVLM)

The slides can be downloaded via this QR code

References

- [Yamada+’21] Kosuke Yamada, Ryohei Sasano, Koichi Takeda: Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering, ACL-IJCNLP 2021 short, pp.811–816
- [Kawahara+’14] Daisuke Kawahara, Daniel Peterson, Octavian Popescu, Martha Palmer: Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses, EACL 2014, pp.58–67
- [Ustalov+’18] Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, Simone Paolo Ponzetto: Unsupervised Semantic Frame Induction using Triclustering, ACL 2018 short, pp.55–62
- [Arefyev+’19] Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, Alexander Panchenko: Neural GRANNy at SemEval-2019 Task 2: A combined approach for better modeling of semantic relationships in semantic frame induction, SemEval 2019, pp.31–38
- [Anwar+’19] Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, Alexander Panchenko: HHMM at SemEval-2019 Task 2: Unsupervised Frame Induction using Contextualized Word Embeddings, SemEval 2019, pp.125–129
- [Ribeiro+’19] Eugénio Ribeiro, Vânia Mendonça, Ricardo Ribeiro, David Martins de Matos, Alberto Sardinha, Ana Lúcia Santos, Luísa Coheur: L2F/INESC-ID at SemEval-2019 Task 2: Unsupervised Lexical Semantic Frame Induction using Contextualized Word Representations, SemEval 2019, pp.130–136
- [Yamada+’23a] Kosuke Yamada, Ryohei Sasano, Koichi Takeda: Semantic Frame Induction with Deep Metric Learning, EACL 2023, pp.1833–1845
- [Yamada+’23b] Kosuke Yamada, Ryohei Sasano, Koichi Takeda: Argument Clustering with Deep Metric Learning for Semantic Frame Induction, ACL 2023 Findings, pp.9356–9364
- [Han+’24] Yi Han, Ryohei Sasano, Koichi Takeda: Definition Generation for Automatically Induced Semantic Frame, In ACL 2024 Findings, pp.11112–11118