IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 March 2023, accepted 2 May 2023, date of publication 8 May 2023, date of current version 15 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274489

RESEARCH ARTICLE

Logic-Based Inference With Phrase Abduction Using Vision-and-Language Models

AKIYOSHI TOMIHARI AND HITOMI YANAKA
Department of Information Science, The University of Tokyo, Tokyo 113-0033, Japan

Corresponding author: Akiyoshi Tomihari (tkaaynakiyoshi@is.s.u-tokyo.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) funded by Precursory Research for Embryonic Science and Technology (PRESTO), Japan, under Grant JPMJPR21C8; and in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Japan, under Grant JP20K19868.

ABSTRACT Recognizing Textual Entailment (RTE) is among the most fundamental tasks in natural language processing applications, such as question answering and machine translation. One of the main challenges in logic-based approaches to this task is the lack of background knowledge. This study proposes a logical inference system with phrasal knowledge by comparing their visual representations based on the intuition that visual representations enable people to judge entailment relations. First, we obtain candidate phrase pairs for phrasal knowledge from logical inference. Second, using a vision-and-language model, we acquire the visual representations of these phrases in the form of images or embedding vectors. Finally, we compare these obtained visual representations to determine whether to inject the knowledge corresponding to the candidate. In addition to simple similarity between phrases, we also consider asymmetric relations when comparing visual representations. Our logical inference system improved accuracy on the SICK dataset compared with a previous logical inference system, SPSA (Selector of Predicates with Shared Arguments). Moreover, our asymmetric evaluation functions using vision-and-language models are effective at capturing the entailment relations of word pairs in HyperLex.

INDEX TERMS Natural language processing, recognizing textual entailment, vision and language.

I. INTRODUCTION

Recognizing Textual Entailment (RTE) [1], also known as Natural Language Inference (NLI), is a key task in natural language processing (NLP) applications such as question answering and machine translation. This task predicts whether a given premise sentence entails a hypothesis sentence. Logic-based approaches [2], [3], [4], [5], [6] and machine learning approaches [7], [8], [9] are the primary approaches to RTE. Logic-based approaches use logical formulas to represent the linguistic meanings of sentences and try to prove entailment relations between formulas. On the other hand, machine learning approaches embed sentences in a vector space and train end-to-end neural models for RTE tasks. Machine learning approaches have achieved high performance in RTE. However, they suffer from low interpretability and explainability, and they have some limitations in what they can do, such as their generalization ability [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

In this study, we focus on interpretable logic-based approaches, which successfully provide semantic representations of sentences with linguistically challenging phenomena such as generalized quantifiers and comparatives. Logic-based approaches tend to achieve high precision (the number of correctly predicted entailment labels divided by the total number of predicted entailment labels) on RTE tasks. However, logical inference systems cannot correctly predict entailment labels when they do not have the background knowledge necessary to prove that a given premise entails a hypothesis. This is a main reason for the low recall (the number of correctly predicted entailment labels divided by the total number of entailment labels in all premise–hypothesis pairs given to a system) of such systems [11]. Previous logical inference systems [12], [13] have attempted to address this problem by using text knowledge databases such as WordNet [14] and injecting axioms as background knowledge during a proof. However, these systems still lack the knowledge necessary for completing the proof, especially phrasal knowledge.

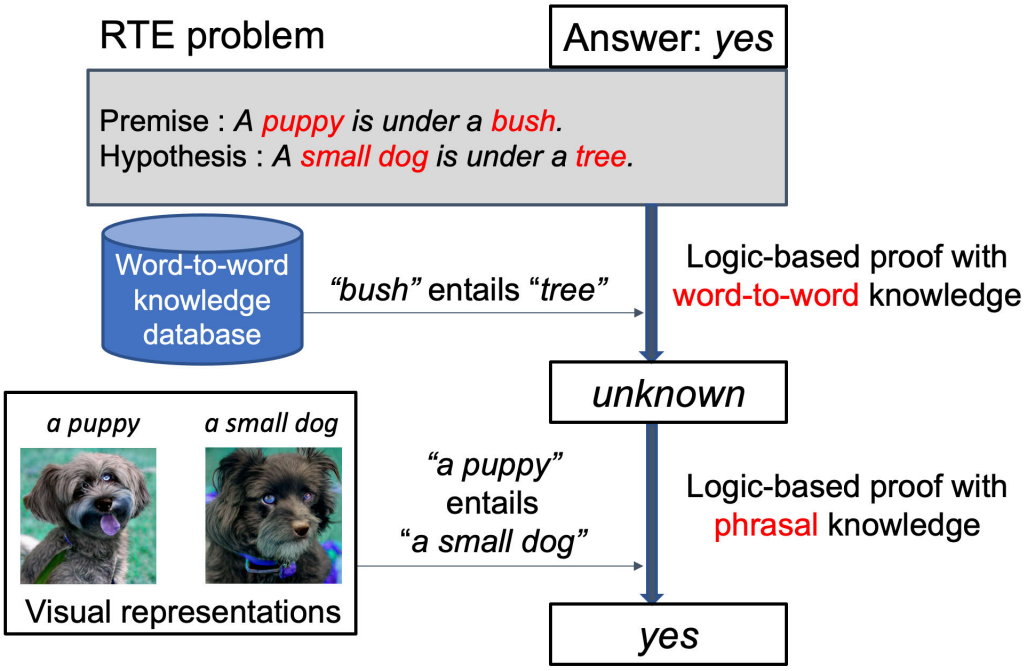


FIGURE 1. The problem setting of RTE tasks with an example. An RTE task determines whether there is an entailment relation between a given premise–hypothesis pair. To solve an RTE task with logic-based approaches, both word-to-word knowledge and phrasal knowledge are needed. There is an entailment relation that the premise *"A puppy is under a bush."* entails the hypothesis *"A small dog is under a tree."* The knowledge that *"bush"* entails *"tree"* can be acquired from word-to-word knowledge databases, but the phrasal knowledge that *"a puppy"* entails *"a small dog"* cannot. Logical inference systems need to acquire this phrasal knowledge to prove the entailment relation. Our system injects phrasal knowledge into the proof using visual representations of vision-and-language models.

Fig. 1 illustrates the problem setting of RTE where phrasal knowledge is necessary for judging entailment relations between texts. In this example, there is an entailment relation that the premise "*A puppy is under a bush.*" entails the hypothesis "*A small dog is under a tree.*" The knowledge that "*bush*" entails "*tree*" can be acquired from word-to-word knowledge databases, but the phrasal knowledge that "*a puppy*" entails "*a small dog*" cannot. Logical inference systems need to acquire this phrasal knowledge to prove the entailment relation.

To overcome the lack of phrasal knowledge, text databases for phrasal knowledge, such as Paraphrase Databases (PPDB) [15], can be used [16], [17] in addition to word-to-word knowledge databases. Furthermore, phrasal knowledge can be acquired using image databases [18] and labeled RTE problems [13], [19]. However, since these approaches depend on knowledge databases, it remains difficult to acquire knowledge of long phrases.

At the same time, recent developments in deep learning have stimulated research in tasks involving both vision and language, such as visual question answering [20], image captioning [21], multimodal machine translation [22], and visual entailment [23]. DALL-E [24] is a text-to-image generation model that automatically generates images from text prompts with high zero-shot performance. CLIP (Contrastive Language-Image Pre-training) [25] is another vision-and-language model that is pre-trained from Web databases to classify images with text labels. The embedding vectors of CLIP have been applied to various vision-and-language tasks [26]. Given the development of vision-and-language models, they can be applied to inject phrasal knowledge into logical inference systems.

In this study, we aim to improve logical inference systems by addressing the lack of phrasal knowledge required to solve RTE problems. To achieve this goal, the main question of our study is as follows: **Can visual representations capture entailment relations between phrases and improve the performance of logical inference systems in RTE tasks?** This hypothesis is based on the intuition that visual representations enable people to judge entailment relations. Given the significant development of vision and language, we use vision-and-language models to acquire visual representations from texts. We define visual representations as images and embedding vectors of vision-and-language models and investigate whether visual representations of phrases support the knowledge from databases needed to judge entailment relations between texts.

In this context, an additional question arises: **How can we effectively evaluate the visual representations of phrases to capture their entailment relations?** Embedding vectors of texts, specifically, are encoded with various semantic and syntactic properties of words, and the challenge lies in effectively extracting these properties. Although cosine similarity is a well-known metric for determining similarity between word embeddings [27], [28], it fails to consider vector norms. Previous studies have shown that norms of word vectors

represent the relative informativeness of words [29], [30], [31], [32]. We investigate whether norms of the embedding vectors of vision-and-language models can be used to capture entailment relations between phrases.

In this study, we aim to improve the performance of ccg2lambda[1] [33], a logical inference system that obtains logical formulas as semantic representations of sentences based on Combinatory Categorial Grammar (CCG) [34] and judges their entailment relations by natural deduction proofs. We extract potential entailment phrase pairs from CCG syntactic trees and semantic representations of input sentences. We acquire visual representations (images or embedding vectors) of these phrase pairs using vision-and-language models, specifically CLIP and DALL-E. By evaluating each phrase pair, we determine whether it has an entailment relation using its visual representations. We propose several evaluation functions for phrase pairs, evaluate our logical inference system on the standard RTE dataset, SICK [35], and evaluate the effectiveness of evaluation functions using embedding vectors of vision-and-language models on graded lexical entailment tasks.

The contributions of this paper are summarized as follows.

- We propose a novel method for injecting phrasal knowledge into a logical inference system using vision-and-language models.
- We provide functions for calculating the degree of entailment relations between phrases from their visual representations generated by vision-and-language models.
- Compared with its baseline system, our logical inference system showed an increase in accuracy on the SICK dataset.
- Our experimental results demonstrate that in a word pair with an entailment relation p and h, and their CLIP embedding vectors $\mathbf{v}_p$ and $\mathbf{v}_h$, the norm of $\mathbf{v}_h$ tends to be greater than that of $\mathbf{v}_p$.

II. RELATED WORK

A. LOGIC-BASED APPROACHES TO RTE TASKS

Logic-based approaches offer interpretability by allowing the pipeline process from input to output entailment relations to be easily understood. They are effective for RTE problems involving a wide range of linguistic phenomena. MartC-nez-GC3mez et al. [33] proposed ccg2lambda, a higher-order inference system that automates natural deduction proofs on compositional semantics of natural language based on CCG [34] parsers and event semantics. Haruta et al. [2], [3] improved the performance of ccg2lambda for problems involving comparatives by combining event semantics [36] and degree semantics [37]. These approaches yielded high precision and succeeded in solving complex RTE problems such as generalized quantifiers and comparatives, which present a challenge to machine learning approaches.

However, a main problem in logic-based approaches to RTE tasks is how to acquire the background knowledge required to solve the problems. To address this issue, MartC-nez-GC3mez et al. [12] added the Selector of Predicates with Shared Arguments (SPSA) mechanism to ccg2lambda. This mechanism injects background knowledge on demand through the use of an interactive natural deduction theorem prover, Coq [38]. Nevertheless, the original SPSA is limited in that it uses only word-to-word knowledge and does not inject phrasal knowledge. Our system supplements SPSA by injecting phrasal knowledge.

Yoshikawa et al. [6] proposed an alternative method for extending ccg2lambda by introducing an efficient mechanism for axiom injection based on Knowledge Base Completion (KBC) models [39]. KBC models are machine learning models that have recently seen significant advancements. Their approach improved the processing speed for solving RTE problems compared with SPSA while maintaining competitive accuracy. However, their method was unable to deal with phrasal knowledge.

There are two problems with injecting phrasal knowledge: one is how to extract arbitrary combinations of phrases, and the other is how to inject phrasal knowledge. Yanaka et al. [13] used natural deduction proofs of RTE problems in ccg2lambda to tackle the first problem. They used subgraph matching with variable unification to extract paraphrases. Abzianidze [19] followed their approach and used the proof processes of a tableau theorem prover. However, both of these approaches were unable to deal with unseen paraphrases.

To solve the second problem with injecting phrasal knowledge, Bjerva et al. [16] and Beltagy et al. [17] both used WordNet and PPDB as lexical knowledge databases. Whereas Bjerva et al. used paraphrase rules in PPDB at the text level before parsing sentences, Beltagy et al. translated these paraphrase rules into logical rules. Han et al. [18] used images to tackle the same problem. They mapped phrases to images and combined their visual features with textual and logic features from a logical inference system, thereby providing an RTE classification model. They used a commercial API to retrieve images from image databases. However, these previous methods depend on databases, which are limited in both quantity and accuracy, especially for phrasal knowledge. We tackle the problem of injecting arbitrary phrasal knowledge by using vision-and-language models.

B. VISION-AND-LANGUAGE MODELS

Recently, there has been a significant surge in research on tasks that involve both computer vision and NLP, referred to as vision-and-language tasks, such as visual question answering [20], image captioning [21], multimodal machine translation [22], and visual entailment [23]. Models designed for these vision-and-language tasks, called vision-and-language models, are required to connect natural language descriptions with their visual representations. Pre-training a large model on large-scale general datasets and then fine-tuning it on specific tasks is a common technique in these areas [40], [41], [42], [43].

[1] https://github.com/mynlp/ccg2lambda

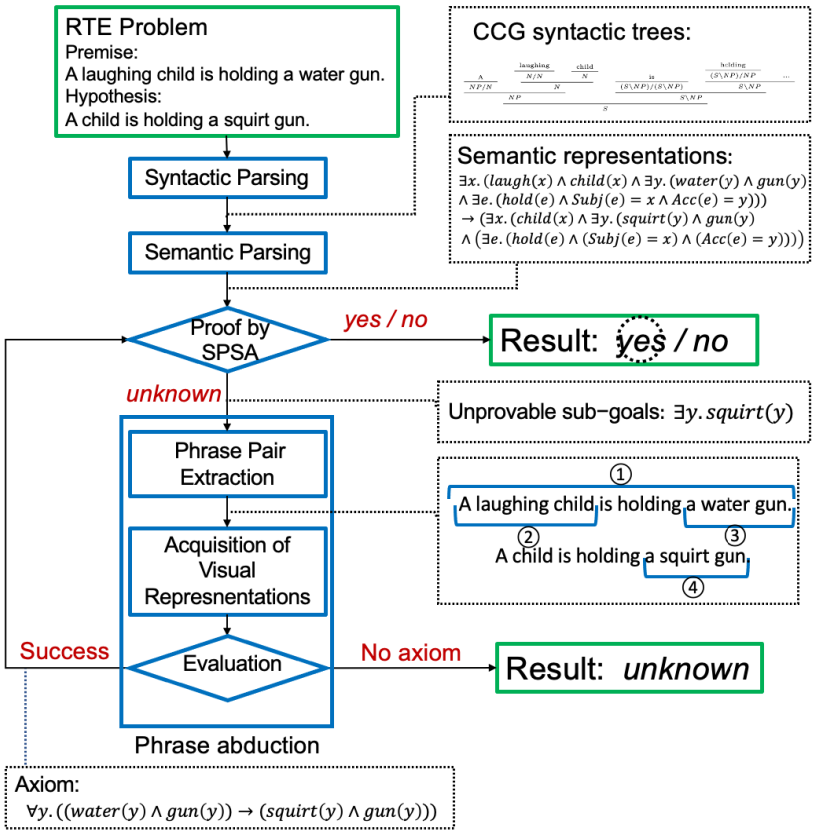


FIGURE 2. **Overview of the proposed method with an example. First, our system transforms a premise and a hypothesis into CCG syntactic trees. It then obtains semantic representations from these syntactic trees. Next, we try to prove whether the premise entails the hypothesis using SPSA (Selector of Predicates with Shared Arguments). If SPSA outputs *unknown*, then we try phrasal axiom injection, which consists of phrase pair extraction, acquisition of the visual representations, and evaluation of the degree of entailment. If we succeed in phrasal axiom injection, then we continue the proof by SPSA with the injected axioms. Otherwise, we finish the proof with the result *unknown*. This process of proof by SPSA and phrasal axiom injection is repeated until either *yes* or *no* is acquired as a result, or until the phrasal axiom injection fails with the result *unknown*.**

One such vision-and-language model, CLIP [25], used both a text encoder and a visual encoder. It was trained on a dataset of 400 million text and image pairs from the Internet which inherently contains a high degree of noise. The text and visual encoders independently encode the input images and texts, respectively. They are pre-trained with a contrastive loss, where the cosine similarity of the image and text embeddings of the real pairs is maximized and that of the incorrect pairs is minimized. As a result, CLIP encodes similar images and texts into similar embedding vectors.

Another notable vision-and-language model is DALL-E [24], which was developed for text-to-image generation tasks. The text encoder of DALL-E is based on an autoregressive transformer [42], and its visual encoder is based on a variational autoencoder [44]. DALL-E was trained on 250 million text and image pairs from the Internet with 12 billion parameters. Both CLIP and DALL-E exhibit high zero-shot performance [26]. We hypothesize that these vision-and-language models are useful for capturing phrasal knowledge in logic-based approaches to RTE tasks.

C. EMBEDDING VECTORS OF TEXTS AND THEIR SIMILARITY

Text embedding vectors were proposed on the premise that texts can be assigned dense, low-dimensional vector representations that capture linguistic relations between them. Some models, such as the skip-gram negative-sampling model [45], rely solely on distributional knowledge derived from textual corpora and bring representations of similar words in proximity. The word frequency ratio model [46] leverages the idea that more general concepts tend to appear more frequently within textual corpora. Santus et al. [47] and Kiela et al. [48] combined symmetric cosine similarity and asymmetric generality measures obtained from texts and visual data, respectively. Gaussian embedding [49] represents words as multivariate Gaussians instead of points in the embedding space, which inherently generates asymmetry.

Other models complement distributional knowledge with external linguistic constraints extracted from WordNet. The POINCARÉ model [50] uses hyperbolic spaces to learn general lexical entailment embeddings based on Poincaré balls

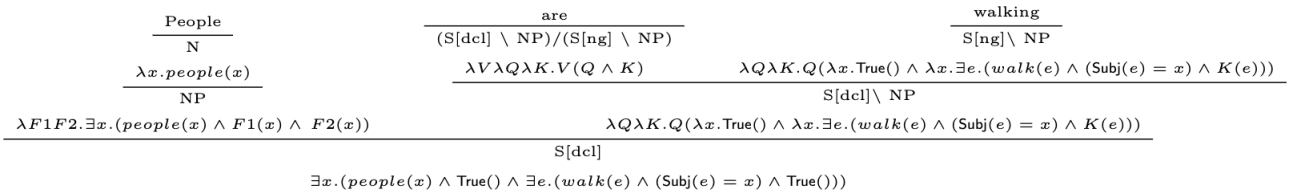


FIGURE 3. Example of a CCG syntactic tree and a semantic representation for the sentence *People are walking*. CCG parsers assign a syntactic category to each lexical item. Our system assigns a meaning, specified as a λ-term, to each leaf of the syntactic tree, which are then combined compositionally according to semantic templates. In our semantic templates, which is based on Neo-Davidson Event Semantics, every verb is decomposed into a one-place predicate over events and a set of functional expressions using auxiliary predicates for semantic roles, such as Subj.

with encoded hierarchy and semantic similarity obtained from WordNet. The LEAR model [51] specializes a word vector space to emphasize asymmetric relations from WordNet by using vector norms, achieving high performance in lexical entailment tasks.

Cosine similarity is commonly used to determine the similarity between word embeddings [27], [28] and is interpreted as capturing the angular information of vectors. Empirical evidence has demonstrated that cosine similarity effectively captures word similarity [52]. It should be noted, however, that cosine similarity disregards vector norms. Moreover, the influence of word frequency information in the embedding space for cosine similarity has been documented [53].

Word vectors have been shown to exhibit additive compositionality, implying that their sum or average within a sentence effectively captures the meaning of that sentence [30], [54]. Additionally, previous studies have shown that the norms of word vectors within a sentence are dispersed [55]. In light of these two observations, it has been suggested that the norm of each word vector serves as a weighting factor in the additive composition of a sentence representation [31]. In the skip-gram negative-sampling model, the squared norm of word embeddings is closely related to the Kullback-Leibler (KL) divergence in the corpus, both theoretically and experimentally [32]. The findings of previous studies have suggested that the norms of word embedding vectors represent the relative informativeness of the words. In this study, we assess whether embedding vectors of vision-and-language models have this property as well.

III. METHODOLOGY

A. SYSTEM OVERVIEW

We consider RTE problems consisting of premise–hypothesis pairs annotated with three relations: entailment (*yes*), contradiction (*no*), and neutral (*unknown*). If the premise is true, entailment (*yes*) indicates that then the hypothesis is also true, while contradiction (*no*) indicates that it cannot be true. Neutral (*unknown*), meanwhile, indicates that whether the hypothesis is true or false is independent of the premise.

Our logical inference system is based on ccg2lambda [33]. We extend SPSA [12], a mechanism of axiom injection for word-to-word knowledge, to perform phrase abduction.

Fig. 2 shows an overview of our system. First, our system transforms a premise (P) and a hypothesis (H) into CCG syntactic trees (see Fig. 3 for an example) through CCG parsers, deriving semantic representations of these texts based on the syntactic trees. We provide the details of this process in Section III-B.

Second, our system uses the theorem proving of Coq [38] with SPSA to judge whether entailment ($P \rightarrow H$) or contradiction ($P \rightarrow \neg H$) holds between the premise and the hypothesis. Coq is an interactive natural deduction theorem prover that is fully automatic with several built-in theorem-proving routines known as tactics. During the proof process, SPSA injects axioms for word-to-word knowledge. Additionally, our system injects axioms for phrasal knowledge using visual representations (images or embedding vectors) generated from a vision-and-language model (DALL-E or CLIP) using the following proof process.

1) Try to prove whether the premise entails the hypothesis using SPSA (Section III-C). If the result is *yes* or *no*, then finish the proof. If not, then proceed to Step 2.
2) Try phrase abduction using a vision-and-language model (DALL-E or CLIP) (Sections III-D1 and III-D2). If at least one phrasal axiom is injected, then return to Step 1. If not, then finish the proof with the result *unknown*.

We provide an example of this proof process in Section III-E.

B. SYNTACTIC AND SEMANTIC PARSING

To start, our system parses the premise (P) and the hypothesis (H) using CCG [34] parsers, which assign a syntactic category to each lexical item. CCG is a lexicalized grammar formalism that provides a transparent interface between syntax and semantics.

In the CCG syntactic trees, our system assigns a meaning, specified as a λ-term, to each leaf of the syntactic tree and combines compositionally according to semantic templates. For semantic templates, we adopt Neo-Davidson Event Semantics [36], in which a sentence is mapped to a formula involving quantification over events. In Neo-Davidson Event Semantics, every verb is decomposed into a one-place predicate over events and a set of functional expressions using auxiliary predicates for semantic roles, such as Subj (see Fig. 3 for examples). This process of syntactic and semantic parsing yields logical formulas for input sentences.

C. WORD ABDUCTION

Our system tries to prove the entailment relation ($P \rightarrow H$) and the contradiction relation ($P \rightarrow \neg H$) using Coq and the

logical formulas of the input sentences. In the proof process of Coq, the logical formulas are decomposed into sets of formulas without logical connectives, which are known as atomic formulas. The goal of the proof is to prove all atomic formulas of the right-hand side of the right arrow (H or $\neg H$), which are referred to as sub-goals.

Coq searches for sub-goals that share the same predicate with atomic formulas of the premise. If any are found, then they are proved and the variables in their arguments are unified.

If unprovable sub-goals remain, then the naive proof by Coq fails. In such cases, SPSA searches for predicates in the premise that share the same arguments with the unprovable sub-goals. If any are found, then SPSA checks the linguistic relations between the word of the premise and that of the sub-goal using two word-to-word knowledge databases, namely, WordNet [14] and VerbOcean [56]. If the linguistic relations are recognized, then the corresponding axioms are provided and injected into the proof process. This process of word axiom injection is called word abduction.

D. PHRASE ABDUCTION

When the result of proof by SPSA is *unknown*, we proceed to Step 2 of Section III-A. To perform phrase abduction, we extract phrase pairs that are candidates for phrasal knowledge and evaluate them by calculating the degree of their entailment relations, as described in Section III-D1. We use visual representations of vision-and-language models in the evaluation step, which is explained in Section III-D2. After the evaluation, the phrase pair with the highest evaluation value is selected for each sub-goal. If the value is higher than a certain threshold (set by preliminary experiments), then we determine the phrase pair as a phrasal axiom. In the absence of any entailment phrase pairs, no phrasal axioms are injected, and the proof finishes with the result *unknown*. If there is at least one entailment phrase pair, then the corresponding axioms are injected into the proof process, and SPSA continues the proof. The phrasal axiom is formulated through the conversion of the corresponding semantic representations of both phrases into a Coq script.Â This is our proposed mechanism of phrase abduction. The process of proof by SPSA and phrase abduction is repeated until the result *yes* or *no* is achieved, or until the phrase abduction fails with the result *unknown*.

1) PHRASE PAIR EXTRACTION

We define a phrase as any part of a sentence corresponding to an NP (Noun Phrase), a VP (Verb Phrase), a PP (Prepositional Phrase), or an S (Sentence) in CCG syntactic trees. When the result of proof by SPSA is *unknown* we extract phrase pairs that are candidates for phrasal knowledge for each of the unprovable sub-goals, using CCG syntactic trees and semantic representations to extract phrase pairs. We extract all phrases from the premise, but we extract only the minimum phrase containing the unprovable sub-goal in the semantic representations from the hypothesis. All pairs of extracted phrases from the premise and hypothesis are regarded as extracted phrase pairs.

2) PHRASE PAIR EVALUATION USING VISUAL REPRESENTATIONS

We define visual representations as both images and embedding vectors. To acquire visual representations of phrase pairs, we use two vision-and-language models, namely, DALL-E [24] and CLIP [25]. DALL-E, a text-to-image generation model, generates images automatically from text prompts. CLIP is pre-trained to classify images and encode similar images and texts into similar embedding vectors.

We generate either

- images from DALL-E or
- embedding vectors from the encoder of DALL-E or CLIP

as visual representations. For each phrase pair, we calculate the degree of their entailment relations using their visual representations. For this calculation, we define symmetric and asymmetric evaluation functions for both embedding vectors and images.

a: IMAGES

Using DALL-E, we generate two images each from the premise and hypothesis of a phrase pair (four images in total). We define two types of evaluation functions, namely, symmetric and asymmetric. The symmetric function is given by

$$\mathrm{Sym}_{\mathrm{Im}}(p,h) = 1 - \frac{\sum_{i_p \in I_p} \sum_{i_h \in I_h} (\mathrm{f}(i_p, i_h) + \mathrm{f}(\mathrm{gray}(i_p), \mathrm{gray}(i_h)))}{Z \cdot |I_p| \cdot |I_h|}, \quad (1)$$

and the asymmetric function by

$$\mathrm{Asym}_{\mathrm{Im}}(p,h) = 1 - \frac{\sum_{i_p \in I_p} \min_{i_h \in I_h} \{\mathrm{f}(i_p, i_h) + \mathrm{f}(\mathrm{gray}(i_p), \mathrm{gray}(i_h))\}}{\cdot |I_p|}, \quad (2)$$

where

- p and h represent the premise and hypothesis of a phrase pair,
- I_p and I_h are the sets of images generated from p and h,
- gray is a function that transforms an image to grayscale,
- Z is a normalization constant, $Z = 3000$, that ensures the return value is in the range of 0 to 1, and
- f is a function from AugNet [57] that calculates the distance between two images.

The symmetric function $\mathrm{Sym}_{\mathrm{Im}}$ calculates the average distance between all pairs of images generated from the two phrases. In contrast, the asymmetric function $\mathrm{Asym}_{\mathrm{Im}}$ calculates the averageÂ distance between the best image correspondences. The design of the asymmetric function follows the work of Han et al. [18]. Our preliminary experiments have

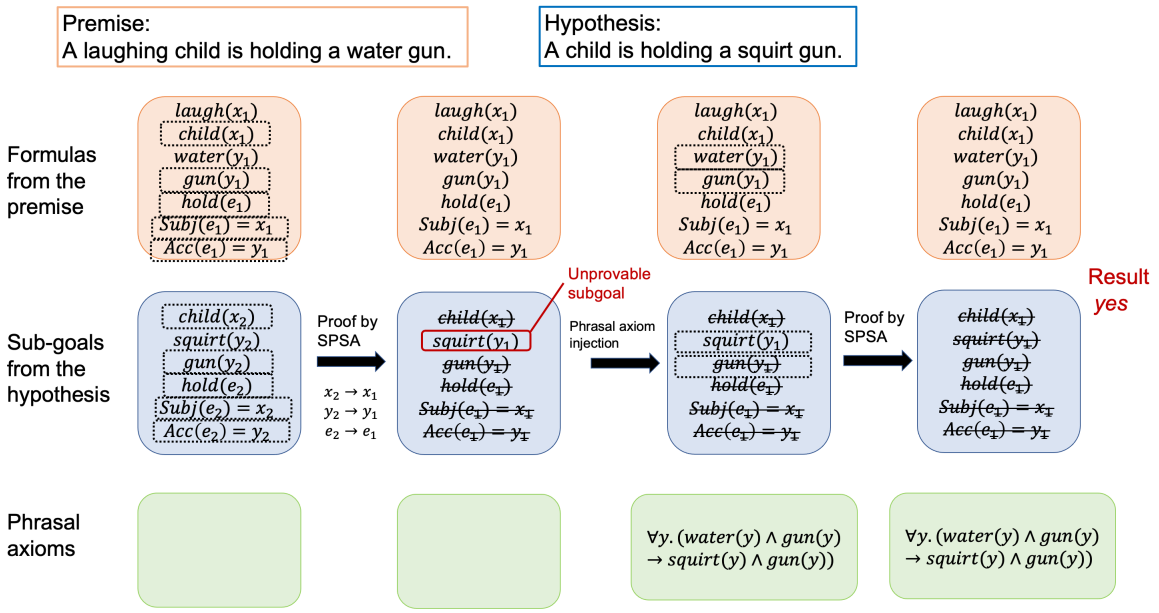


FIGURE 4. Example of the proof process of our logical inference system. First, SPSA tries to prove sub-goals from the hypothesis in the blue box using predicates in the orange box. It finds common predicates from the premise and hypothesis, proves them in the hypothesis, and unifies the variables. SPSA also tries to find a predicate that shares an argument with unprovable sub-goal $squirt(y_1)$ and entails the word "*squirt*" using two word-to-word knowledge databases, namely, WordNet and VerbOcean. However, such a premise cannot be found, so SPSA does not inject anyÂ axiom and the result is *unknown*. Second, our system tries phrase abduction using visual representations (images or embedding vectors) and the axiom $\forall y.(water(y) \wedge gun(y) \rightarrow squirt(y) \wedge gun(y))$ is injected. Third, SPSA once again tries to prove the unprovable sub-goal $squirt(y_1)$ with the injected axiom. SPSA succeeds this time, so SPSA finishes the proof with the result *yes*.

shown that using grayscale images in combination with color images leads to better performance than using color images alone. Therefore, we take the average distance between the original images $\mathrm{f}(i_p, i_h)$ and the distance between grayscale images $\mathrm{f}(\mathrm{gray}(i_p), \mathrm{gray}(i_h))$ as the distance between images i_p and i_h.

b: EMBEDDING VECTORS

Using the encoder of DALL-E or CLIP, we generate embedding vectors for both the premise and hypothesis of a phrase pair. To evaluate phrase pairs, we define three types of evaluation functions, a symmetric function, an asymmetric function, and their product, which are given by the following expressions:

$$\mathrm{Sym}_{\mathrm{Em}}(p,h) = \cos(\mathbf{v}_p, \mathbf{v}_h) = \frac{(\mathbf{v}_p \cdot \mathbf{v}_h)}{|\mathbf{v}_p| \cdot |\mathbf{v}_h|} \tag{3}$$

$$\mathrm{Asym}_{\mathrm{Em}}(p,h) = \frac{|\mathbf{v}_h| - |\mathbf{v}_p|}{(|\mathbf{v}_p| + |\mathbf{v}_h|)/2} \tag{4}$$

$$\begin{aligned}\mathrm{Mul}_{\mathrm{Em}}(p,h) &= \mathrm{Sym}_{\mathrm{Em}}(p,h) \cdot \mathrm{Asym}_{\mathrm{Em}}(p,h) \\ &= \frac{(\mathbf{v}_p \cdot \mathbf{v}_h)}{|\mathbf{v}_p| \cdot |\mathbf{v}_h|} \cdot \frac{|\mathbf{v}_h| - |\mathbf{v}_p|}{(|\mathbf{v}_p| + |\mathbf{v}_h|)/2}\end{aligned} \tag{5}$$

where

- p and h represent the premise and hypothesis of a phrase pair,
- $\mathbf{v}_p$ and $\mathbf{v}_h$ are the embedding vector of phrases p and h, respectively, and
- cos is the function used to calculate the cosine similarity between two vectors.

The symmetric function $\mathrm{Sym}_{\mathrm{Em}}$ gives the cosine similarity between $\mathbf{v}_p$ and $\mathbf{v}_h$. The asymmetric function $\mathrm{Asym}_{\mathrm{Em}}$ is designed based on the approach proposed by Vulic et al. [51]. It gives the difference between the norms of $\mathbf{v}_h$ and $\mathbf{v}_p$, then normalizes the result by dividing it by the mean of the norms. Finally, $\mathrm{Mul}_{\mathrm{Em}}$ is obtained by multiplying the values of these two functions.

E. EXAMPLE OF THE PROOF PROCESS

Fig. 4 shows an example of the proof process used by our system with an input RTE problem with the premise "*A laughing child is holding a water gun.*" and the hypothesis "*A child is holding a squirt gun.*" In this example, there is an entailment relation between the premise (P) and the hypothesis (H), and our system tries to prove this relation ($P \rightarrow H$).

The logical formulas for the premise (P) and the hypothesis (H), which are acquired through syntactic and semantic parsing, are decomposed into the atomic formulas in the orange and blue boxes, respectively. The goal of the proof is to prove all of the sub-goals in the blue box.

First, in Step 1 of Section III-A, the proof is performed through SPSA. Coq attempts to prove sub-goals in the blue box and to prove sub-goals that share the same predicate with atomic formulas of the premise in the orange box. As a result, $child(x_2)$, $gun(x_2)$, $hold(x_2)$, $Subj(e_2) = x_2$, and $Acc(e_2) = y_2$ are proved with an unproved sub-goal $squirt(y_1)$. Accordingly, x_2, y_2, and e_2 are unified to x_1, y_1, and e_1, respectively. SPSA tries to find logical formulas in the premise that have

TABLE 1. Examples of RTE tasks in the SICK dataset.

Example	Entailment
Premise: *A man, a woman, and two girls are walking on the beach.* Hypothesis: *A group of people is near the sea.*	*yes* (entailment)
Premise: *Two dogs are wrestling and hugging.* Hypothesis: *There is no dog wrestling and hugging.*	*no* (contradiction)
Premise: *A girl has a t-shirt and is keeping her mouth open.* Hypothesis: *A girl from Asia, in front of a brick window, looks surprised.*	*unknown* (neutral)

TABLE 2. Results of RTE tasks on the SICK dataset. VR and EF denote visual representation and evaluation function, respectively.

VR	Model	EF	Acc.	Prec.	Rec.	F-measure
Image	DALL-E	Sym_{Im}	83.26	96.76	63.03	76.33
		Asym_{Im}.	83.24	78.50	62.98	76.30
Embedding Vector	DALL-E	Sym_{Em}	83.17	96.82	62.79	76.18
		Asym_{Em}	82.53	93.04	63.64	75.59
		Mul_{Em}	81.63	89.47	64.48	74.95
	CLIP	Sym_{Em}	83.36	93.97	64.95	76.81
		Asym_{Em}	82.44	92.28	64.39	75.85
		Mul_{Em}	83.13	96.28	63.07	76.22
SPSA			83.13	96.88	62.65	76.08
Han et al. (2017)			84.29	90.24	71.08	79.52
No axioms			76.65	98.90	46.48	63.24

the argument y_1, and $water(y_1)$ and $gun(y_1)$ are identified as such formulas. However, the linguistic relations between *water* and *squirt*, or *gun* and *squirt*, are not obtained from the word-to-word databases. Consequently, no axiom is injected by word abduction, and the proof is finished with the result *unknown*.

Next is Step 2, in which our system attempts phrase abduction using a vision-and-language model (DALL-E or CLIP). From the premise, the phrases ①"*a laughing child is holding a water gun*" (S), ②"*a laughing child*" (NP), and ③"*a water gun*" (NP) are extracted. From the hypothesis, only ④"*a squirt gun*" (NP) is extracted, which is the minimum phrase that contains the unprovable sub-goal $\exists y_1.squirt(y_1)$ (see Fig. 2). All pairs of these phrases extracted from the premise and hypothesis, namely ("*a laughing child is holding a water gun*", "*a squirt gun*"), ("*a laughing child*", "*a squirt gun*"), and ("*a water gun*", "*a squirt gun*") are regarded as extracted phrase pairs. In the evaluation step, the phrase pair ("*a water gun*", "*a squirt gun*") gets the highest evaluation, and the evaluation value exceeds the threshold. Therefore, the corresponding axiom $\forall y.((water(y) \wedge gun(y)) \rightarrow (squirt(y) \wedge gun(y)))$ is injected, and our system returns to Step 1.

Finally, SPSA once again tries to prove the unprovable sub-goal $squirt(y_1)$ with the injected axiom. This time it succeeds, so the proof is finished with the result *yes*.

IV. EXPERIMENT I: RTE

A. DATASET

We used the SemEval-2014 version of the SICK dataset [35] to evaluate our logical inference system on RTE tasks. The SICK dataset is a source of English single-premise RTE problems. This dataset was originally developed to evaluate approaches of compositional distributional semantics, and it contains problems involving various lexical, syntactic, and semantic phenomena. Therefore, solving problems in this dataset requires both lexical and phrasal knowledge. This dataset has been used to assess the performance of logical inference systems, such as SPSA. It contains problems with train/trial/test splits of 4500/500/4927 premise–hypothesis pairs and an entailment(*yes*)/contradiction(*no*)/neutral(*unknown*) label distribution of 0.29/0.15/0.56. The SICK dataset comprises 2,409 different words, with an average sentence length of 10.6 words (see Table 1 for examples). We evaluated our system on the test set.

B. EXPERIMENTAL SETUP

We used three CCG parsers, C&C [58], EasyCCG [59], and depccg [60], to mitigate parsing errors. In the case where one result is *yes* and another *no*, our system outputs *unknown*. In other cases, if at least one parser results in *yes* or *no*, then our system outputs the result. The thresholds used in the evaluation step (III-D2) were determined based on the accuracy of 300 problems sampled from the train set of the SICK dataset. We compared our results with those of SPSA [12], an RTE classification model trained with images from dataset [18], and the system without axiom injection (no axiom).

C. OVERALL RESULTS

Table 2 shows the results of logical inference systems on the SICK dataset. Our system outperformed the baseline system, SPSA, on nearly all the experimental conditions. The highest accuracy was achieved on the condition CLIP–Embedding Vector–Sym_{Em} (83.36%) without any training. When we used DALL-E as a vision-and-language model, using images as visual representations rather than embedding vectors yielded higher accuracy (83.26% vs. 83.17%).

D. ANALYSIS

Table 3 shows examples of the RTE results on the SICK dataset, for which the result of our baseline system, SPSA, was *unknown*. Fig. 5 shows images generated by DALL-E when our system evaluated the phrase pairs in the problems of these examples. ID 4589 is a successful example, which

TABLE 3. Examples of the RTE results on the SICK dataset, for which the result of SPSA was *unknown*. We changed the following three conditions of the experiment: (1) model: DALL-E or CLIP, (2) visual representation: Image or Embedding Vector, and (3) evaluation function: Sym_{Im}, Asym_{Im} (for Image), Sym_{Em}, Asym_{Em}, or Mul_{Em} (for Embedding Vector). The answer can be either *yes* (Y, entailment), *no* (N, contradiction), or *unknown* (U, neutral). If the result of a system matched the gold answer, then it is marked with ✓. If not, then it is marked with ✗.

				DALL-E					CLIP		
				Image		Embedding Vector			Embedding Vector		
ID	Premise	Hypothesis	Ans	Sym_{Im}	Asym_{Im}	Sym_{Em}	Asym_{Em}	Mul_{Em}	Sym_{Em}	Asym_{Em}	Mul_{Em}
4589	*A hamster is singing.*	*A small animal is singing.*	Y	✓	✓	✓	✗(U)	✓	✓	✓	✓
3184	*A man is trekking in the woods.*	*The man is hiking in the woods.*	Y	✓	✓	✓	✗(U)	✗(U)	✓	✗(U)	✗(U)
6853	*A dog is running downhill.*	*A dog is running uphill.*	U	✗(Y)	✗(Y)	✗(Y)	✓	✗(Y)	✗(Y)	✓	✓
558	*A dark black dog and a light brown dog are playing in the backyard.*	*A dark black dog and a light brown dog are fighting in the backyard.*	U	✗(Y)	✗(Y)	✗(Y)	✓	✓	✗(Y)	✓	✓

(a) *a hamster* (c) *a man is trekking in the woods* (e) *a dog is running downhill* (g) *a dark black dog and a light brown dog are playing in the backyard*

(b) *a small animal* (d) *the man is hiking in the woods* (f) *a dog is running uphill* (h) *a dark black dog and a light brown dog are fighting in the backyard*

FIGURE 5. Images generated by DALL-E from phrase pairs in the SICK dataset. The phrase pairs of ID 4589 ((a) and (b)) and ID 3184 ((c) and (d)) are true phrasal knowledge and those of ID 6853 ((e) and (f)) and ID 558 ((g) and (h)) are false phrasal knowledge. The problems and the results of the experiment are presented in Table 3.

requires recognizing the phrasal knowledge that "*a hamster*" entails "*a small animal*". Although SPSA could not recognize this entailment phrase pair, all of our systems with DALL-E or CLIP, except for the condition CLIP–Embedding Vector–Asym_{Em}, could recognize this entailment phrase pair. As a result, the corresponding axiom $\forall x.\mathit{hamster}(x) \rightarrow (\mathit{small}(x) \wedge \mathit{animal}(x))$ was injected, and the system proved the correct result *yes*. Fig. 5 (a) and (b) show that similar images were generated from the phrase pair.

The phrase pair for ID 3184 represents true phrasal knowledge that "*a man is trekking in the woods*" entails "*the man is hiking in the woods*". Fig. 5 (c) and (d) show that similar images were generated from these two phrases. However, when we used the embedding vectors and asymmetric functions, this phrase pair was mistakenly not recognized as phrasal knowledge, and the result was *unknown*. This suggests that asymmetric functions cannot recognize entailment phrase pairs that are semantically similar and that symmetric functions are adequate for injecting paraphrases.

Both phrase pairs for problem IDs 6853 and 558 represent false phrasal knowledge that "*a dog is running downhill*" entails "*a dog is running uphill*" and "*a dark dog and a light brown dog are playing in the backyard*" entails "*a dark dog and a light brown dog are fighting in the backyard*." However, similar images were generated from these two phrases, shown in Fig. 5 (e), (f), (g), and (h), and the phrasal knowledge was injected when we used either images or both embedding vectors and symmetric functions. In contrast, this false phrasal knowledge was not injected when we used both embedding vectors and asymmetric functions. This suggests

TABLE 4. Examples in Hyperlex. HyperLex is a standard dataset for evaluating how well word representation models capture graded lexical entailment.

Word class	Word pairs (X, Y)	Score
Nouns	(reason, cause)	5.77
	(worker, waitress)	1.00
Verbs	(stew, cook)	5.57
	(perform, sing)	2.64

that asymmetric functions contributed to reducing the over-generation of axioms in some cases.

V. EXPERIMENT II: GRADED LEXICAL ENTAILMENT TASK

Capturing the entailment relations of texts is crucial for injecting phrasal knowledge into logical inference systems. To investigate whether vision-and-language models can capture entailment relations of phrases, we performed a detailed analysis of embedding vectors of vision-and-language models on graded lexical entailment tasks. We tested the effectiveness of our evaluation functions Sym_{Em}, Asym_{Em}, and Mul_{Em} in calculating word pairs' degree of entailment (known as a graded lexical entailment task) on HyperLex [61].

A. DATASET

We used HyperLex [61], which is a standard dataset for evaluating how well word representation models capture graded lexical entailment. This dataset is grounded in the notions of concept typicality [62] and category vaguness [63] from cognitive science. HyperLex contains 2,616 word pairs (2,163 noun pairs and 453 verb pairs) scored from 0 to 6 by human raters for the following question: "To what degree is X a type of Y?" (see Table 4 for examples). Note that HyperLex contains only word pairs and does not contain a phrase pair.

B. EXPERIMENTAL SETUP

For each word pair, we encoded these words into the embedding vectors of DALL-E and CLIP. Using these embedding vectors, we scored their degree of entailment with our evaluation functions Sym_{Em}, Asym_{Em}, and Mul_{Em}. We scored all 2,616 word pairs from HyperLex and computed Spearman's rank correlation with the ground-truth ranking, following Nickel et al. [50].

C. RESULTS AND ANALYSIS

Table 5 shows performances for the graded lexical entailment task on HyperLex. Vision-and-language models are compared with distributional representation models that rely solely on distributional knowledge derived from textual corpora, which are evaluated by Vulić et al. [61] (the middle 5 rows: **FREQ-RATIO** [46], **SGNS** [45], **SLQS-SIM** [47], **VISUAL** [48], and **WORD2GAUSS** [49]) and two recent architectures (the bottom 2 rows: **POINCARÉ** [50] and **LEAR** [51]), which complement the distributional knowledge with external linguistic constraints derived from WordNet. The details of these distributional representation models are given in Section II-C.

TABLE 5. Spearman's rank correlation scores for the graded lexical entailment task on HyperLex. The top two rows are the results for the vision-and-language models (DALL-E and CLIP) with our evaluation functions. The middle five rows are distributional models that contain only distributional knowledge from textual corpora, which were evaluated by Vulić et al. The bottom two rows are recent architectures, which focus on lexical entailment and complement distributional knowledge with external linguistic constraints extracted from WordNet.

Model	Measure	Result
DALL-E	Sym_{Em}	-0.029
	Asym_{Em}	0.001
	Mul_{Em}	0.001
CLIP	Sym_{Em}	0.145
	Asym_{Em}	0.298
	Mul_{Em}	0.300
FREQ-RATIO		0.279
SGNS		0.205
SLQS-SIM		0.228
VISUAL		0.209
WORD2GAUSS		0.206
POINCARÉ (Nouns)		0.512
LEAR		0.686

These results show that the scores calculated from the embedding vectors of DALL-E do not reflect the degree of lexical entailment, whereas those calculated from embedding vectors of CLIP show a correlation. CLIP is outperformed by **POINCARÉ** and **LEAR** models, which focus on lexical entailment, but it still outperforms other distributional representation models.

This suggests that CLIP embedding vectors with the asymmetric function Asym_{Em} and Mul_{Em} are capable of capturing directional lexical entailment relations. The values of $\text{Asym}_{\text{Em}}(p, h)$ and $\text{Mul}_{\text{Em}}(p, h)$ become positive when the embedding vector norm of the hypothesis phrase h is greater than that of the premise phrase p. Therefore, the results imply that in a word pair with an entailment relation p and h, and their CLIP embedding vectors $\mathbf{v}_p$ and $\mathbf{v}_h$, the norm of $\mathbf{v}_h$ tends to be greater than that of $\mathbf{v}_p$. This differs from the property of distributional representation models where the norm of an embedding vector represents the relative importance of the word [31], [32].

VI. CONCLUSION

In this paper, we presented a novel approach for addressing the lack of phrasal knowledge in RTE tasks by introducing a phrase abduction mechanism of logical inference systems. Our mechanism uses visual representations generated from the vision-and-language models DALL-E and CLIP.

For phrase pairs, we used two types of visual representations (embedding vectors and images) and two types of evaluation functions (symmetric and asymmetric).

Our inference system improved the accuracy of ccg2lambda on the SICK dataset compared to a previous system that used the word abduction mechanism SPSA. Additionally, the results of the graded lexical entailment task on HyperLex suggest that the asymmetric functions we calculated from the embedding vectors of the CLIP model capture the degree of lexical entailment relations more accurately than those of previous distributional representation models without external linguistic constraints. With visual information, our inference system can be applied in a wide range of domains such as object recognition, robotics, and autonomous vehicles to ensure the reliability of operations.

However, there is phrasal knowledge that our symmetric and asymmetric evaluation functions are unable to recognize. Designing better evaluation functions is a task that we will undertake in future work, as is determining how best to complement information that is difficult to represent visually.

ACKNOWLEDGMENT

The authors would like to thank Tomoya Kurosawa, Tomoki Sugimoto, and the other laboratory members for their helpful advice when writing their article.

REFERENCES

[1] I. Dagan, D. Roth, M. Sammons, and F. M. Zanzotto, *Recognizing Textual Entailment*. Cham, Switzerland: Springer, 2013.

[2] I. Haruta, K. Mineshima, and D. Bekki, "Combining event semantics and degree semantics for natural language inference," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, Dec. 2020, pp. 1758–1764.

[3] I. Haruta, K. Mineshima, and D. Bekki, "Logical inferences with comparatives and generalized quantifiers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2020, pp. 263–270.

[4] H. Hu, Q. Chen, K. Richardson, A. Mukherjee, L. S. Moss, and S. Kuebler, "MonaLog: A lightweight system for natural language inference based on monotonicity," in *Proc. Soc. Comput. Linguistics (SCiL)*, New York, NY, USA, 2020, pp. 334–344.

[5] K. Mineshima, P. Martínez-Gómez, Y. Miyao, and D. Bekki, "Higher-order logical inference with compositional semantics," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Lisbon, Portugal, 2015, pp. 2055–2061.

[6] M. Yoshikawa, K. Mineshima, H. Noji, and D. Bekki, "Combining axiom injection and knowledge base completion for efficient natural language inference," in *Proc. AAAI*, Honolulu, HI, USA, 2019, pp. 7410–7417.

[7] A. Lai and J. Hockenmaier, "Illinois-LH: A denotational and distributional approach to semantics," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, 2014, pp. 329–334.

[8] S. Wang and J. Jiang, "Learning natural language inference with LSTM," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, San Diego, CA, USA, 2016, pp. 1442–1451.

[9] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei, "Neural natural language inference models enhanced with external knowledge," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 2406–2417.

[10] M. Glockner, V. Shwartz, and Y. Goldberg, "Breaking NLI systems with sentences that require simple lexical inferences," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 650–655.

[11] J. Bos, "Is there a place for logic in recognizing textual entailment," *Linguistic Issues Lang. Technol.*, vol. 9, no. 3, pp. 1–18, 2014. [Online]. Available: https://aclanthology.org/2014.lilt-9.3

[12] P. Martínez-Gómez, K. Mineshima, Y. Miyao, and D. Bekki, "On-demand injection of lexical knowledge for recognising textual entailment," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Valencia, Spain, 2017, pp. 710–720.

[13] H. Yanaka, K. Mineshima, P. Martínez-Gómez, and D. Bekki, "Acquisition of phrase correspondences using natural deduction proofs," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*. New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 756–766.

[14] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[15] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB: The paraphrase database," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Atlanta, GA, USA, 2013, pp. 758–764.

[16] J. Bjerva, J. Bos, R. van der Goot, and M. Nissim, "The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, Aug. 2014, pp. 642–646.

[17] I. Beltagy, S. Roller, P. Cheng, K. Erk, and R. J. Mooney, "Representing meaning with a combination of logical and distributional models," *Comput. Linguistics*, vol. 42, no. 4, pp. 763–808, Dec. 2016. [Online]. Available: https://aclanthology.org/J16-4007

[18] D. Han, P. Martínez-Gómez, and K. Mineshima, "Visual denotations for recognizing textual entailment," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Copenhagen, Denmark, 2017, pp. 2853–2859.

[19] L. Abzianidze, "Learning as abduction: Trainable natural logic theorem prover for natural language inference," in *Proc. 9th Joint Conf. Lexical Comput. Semantics (SEM)*. Barcelona, Spain: Association for Computational Linguistics, Dec. 2020, pp. 20–31.

[20] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[21] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Feb. 2019.

[22] T. Hirasawa, M. Kaneko, A. Imankulova, and M. Komachi, "Pre-trained word embedding and language model improve multimodal machine translation: A case study in Multi30K," *IEEE Access*, vol. 10, pp. 67653–67668, 2022.

[23] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *CoRR*, vol. abs/1901.06706, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1901.06706

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Lear. (ICML)*, M. Meila and T. Zhang, Eds., vol. 139, 2021, pp. 8821–8831.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Lear. (ICML)*, vol. 139, 2021, pp. 8748–8763.

[26] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How much can CLIP benefit vision-and-language tasks?" 2021, *arXiv:2107.06383*.

[27] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

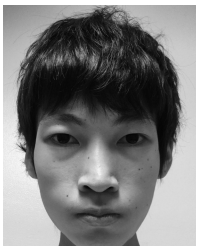
[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

[29] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

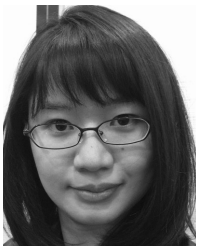
[30] S. Yokoi, R. Takahashi, R. Akama, J. Suzuki, and K. Inui, "Word rotator's distance," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 2944–2960. [Online]. Available: https://aclanthology.org/2020.emnlp-main.236

[31] A. M. J. Schakel and B. J. Wilson, "Measuring word significance using distributed representations of words," 2015, *arXiv:1508.02297*.

[32] M. Oyama, S. Yokoi, and H. Shimodaira, "Norm of word embedding encodes information gain," 2022, *arXiv:2212.09663*.
[33] P. Martínez-Gómez, K. Mineshima, Y. Miyao, and D. Bekki, "ccg2lambda: A compositional semantics system," in *Proc. Assoc. Comput. Linguistics Syst. Demonstrations*, 2016, pp. 1–6.
[34] M. Steedman, *The Syntactic Process*. Cambridge, MA, USA: MIT Press, 2000.
[35] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 216–223.
[36] T. Parsons, *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, MA, USA: MIT Press, 1994.
[37] M. J. Cresswell, "The semantics of degree," in *Montague Grammar*. Amsterdam, The Netherlands: Elsevier, 1976, pp. 261–292.
[38] Y. Bertot and P. Castéran, *Interactive Theorem Proving and Program Development*. Berlin, Germany: Springer, 2004.
[39] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Red Hook, NY, USA: Curran Associates, 2013.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[43] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 5436–5443.
[44] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., Banff, AB, Canada, 2014, pp. 14–16.
[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, Scottsdale, AZ, USA, 2013.
[46] D. Kiela, F. Hill, and S. Clark, "Specializing word embeddings for similarity or relatedness," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Lisbon, Portugal, Sep. 2015, pp. 2044–2048.
[47] E. Santus, A. Lenci, Q. Lu, and S. S. I. Walde, "Chasing hypernyms in vector spaces with entropy," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Gothenburg, Sweden, Apr. 2014, pp. 38–42.
[48] D. Kiela, L. Rimell, I. Vulić, and S. Clark, "Exploiting image generality for lexical entailment detection," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP)*, Beijing, China, Jul. 2015, pp. 119–124.
[49] L. Vilnis and A. McCallum, "Word representations via Gaussian embedding," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.
[50] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6338–6347.
[51] I. Vulić and N. Mrkšić, "Specialising word vectors for lexical entailment," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, New Orleans, LA, USA, 2018, pp. 1134–1145.
[52] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017. [Online]. Available: https://aclanthology.org/Q17-1010
[53] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Lisbon, Portugal, Sep. 2015, pp. 298–307.
[54] J. Mitchell and M. Lapata, "Composition in distributional models of semantics," *Cognit. Sci.*, vol. 34, no. 8, pp. 1388–1429, 2010.
[55] N. Arefyev, P. Ermolaev, and A. Panchenko, "How much does a word weigh? Weighting word embeddings for word sense induction," *CoRR*, vol. abs/1805.09209, Aug. 2018. [Online]. Available: http://arxiv.org/abs/1805.09209
[56] T. Chklovski and P. Pantel, "VerbOcean: Mining the web for fine-grained semantic verb relations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Barcelona, Spain, Jul. 2004, pp. 33–40.
[57] M. Chen, Z. Chang, H. Lu, B. Yang, Z. Li, L. Guo, and Z. Wang, "AugNet: End-to-end unsupervised visual representation learning with image augmentation," *CoRR*, vol. abs/2106.06250, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.06250
[58] S. Clark and J. R. Curran, "Wide-coverage efficient statistical parsing with CCG and log-linear models," *Comput. Linguistics*, vol. 33, no. 4, pp. 493–552, Dec. 2007.
[59] M. Lewis and M. Steedman, "A* CCG parsing with a supertag-factored model," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 990–1000.
[60] M. Yoshikawa, H. Noji, and Y. Matsumoto, "A* CCG parsing with a supertag and dependency factored model," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, 2017, pp. 277–287.
[61] I. Vulić, D. Gerz, D. Kiela, F. Hill, and A. Korhonen, "HyperLex: A large-scale evaluation of graded lexical entailment," *Comput. Linguistics*, vol. 43, no. 4, pp. 781–835, Dec. 2017.
[62] D. L. Medin, M. W. Altom, and T. D. Murphy, "Given versus induced category representations: Use of prototype and exemplar information in classification," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 10, no. 3, pp. 333–352, 1984.
[63] J. A. Hampton, "Typicality, graded membership, and vagueness," *Cognit. Sci.*, vol. 31, no. 3, pp. 355–384, May 2007.



AKIYOSHI TOMIHARI was born in Kawasaki-shi, Kanagawa, Japan, in 2001. He is currently pursuing the B.S. degree in information science with The University of Tokyo, Tokyo, Japan. His current research interests include natural language processing and machine learning.



HITOMI YANAKA received the Ph.D. degree in engineering from The University of Tokyo, Japan, in 2018. From 2018 to 2021, she was a Postdoctoral Researcher with the Riken Center for Advanced Intelligence Project (AIP). She is currently a tenured Lecturer (an Excellent Young Researcher) with the Graduate School of Information Science and Technology, The University of Tokyo. She is also a Visiting Researcher with the Riken AIP. She has also been organizing the International Workshop Natural Logic Meets Machine Learning (NALOMA), since 2020. Her research interests include computational linguistics and natural language processing, with focus on natural language inference.