Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5846–5854
Marseille, 11–16 May 2020

Word Attribute Prediction Enhanced by Lexical Entailment Tasks

Mika Hasegawa, Tetsunori Kobayashi, Yoshihiko Hayashi
Faculty of Science and Engineering, Waseda University
Waseda-machi 27, Shinjuku, Tokyo 169-0042, Japan
mika@pcl.cs.waseda.ac.jp, koba@waseda.jp, yshk.hayashi@aoni.waseda.jp

Abstract

Human semantic knowledge about concepts acquired through perceptual inputs and daily experiences can be expressed as a bundle of attributes. Unlike the conventional distributed word representations that are purely induced from a text corpus, a semantic attribute is associated with a designated dimension in attribute-based vector representations. Thus, semantic attribute vectors can effectively capture the commonalities and differences among concepts. However, as semantic attributes have been generally created by psychological experimental settings involving human annotators, an automatic method to create or extend such resources is highly demanded in terms of language resource development and maintenance. This study proposes a two-stage neural network architecture, Word2Attr, in which initially acquired attribute representations are then fine-tuned by employing supervised lexical entailment tasks. The quantitative empirical results demonstrated that the fine-tuning was indeed effective in improving the performances of semantic/visual similarity/relatedness evaluation tasks. Although the qualitative analysis confirmed that the proposed method could often discover valid but not-yet human-annotated attributes, they also exposed future issues to be worked: we should refine the inventory of semantic attributes that currently relies on an existing dataset.

Keywords: word attributes, word attribute prediction, lexical entailment, similarity, fine-tuning

1. Introduction

A semantic attribute of a concept, such as "an `apple` is *red*", explicitly dictates a semantic aspect of the concept. Thus, given an appropriate set of attributes, a concept can be represented by a bundle of attributes. This notion of semantic attribute (often referred to as semantic feature or semantic property) has some connections with the *componential analysis* in linguistics. A trait of the semantic attributes that they are associated with human percepts and experiences is solidly emphasized by psychological theories, in particular by the notion of *grounded cognition* (Barsalou, 2008).

Unlike the conventional distributed representations (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) purely induced from a text corpus, a semantic attribute is linked with a designated dimension in the vector representations, making their representation appropriate for the computation of commonalities and differences among concepts (Lazaridou et al., 2016; Krebs et al., 2018). This good property could lead to improve the interpretability of an attribute-based NLP system. Moreover, if the semantic attributes even for an unseen concept can be adequately predicted by employing auxiliary information, these attributes can be effectively utilized in several semantic tasks including zero-shot learning/recognition (Al-Halah et al., 2016; Li et al., 2018). In this regard, visual semantic attributes, proven effective in the area of computer vision and the application systems (Lampert et al., 2014; Feris et al., 2017), could provide useful auxiliary information.

However, as semantic attributes are generally created by psychological experimental settings involving human annotators (McRae et al., 2005), an automatic method to create or extend such resources is highly demanded. So far, only a few studies try to map linguistic-based representations to attribute-based representations (Făgărăşan et al., 2015; Bulat et al., 2016). Moreover, the efficacy of these methods is limited, and they pose the issues of coverage and completeness, which are innately associated with human-annotated data.

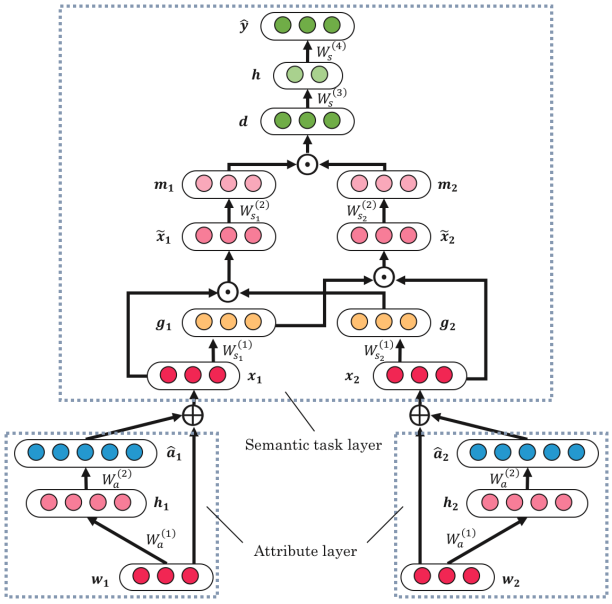

Figure 1: Word2Attr network architecture.

In this study, we propose a two-stage neural network architecture, Word2Attr (Figure 1), to predict the better attribute representation even for an unseen word (Section 3). The proposed network consists of two components: (1) The Attribute Layers, implemented by a multilayer perceptron (MLP), initially learn the mapping from word embedding vectors to their corresponding attribute vectors presented by a human-generated visual attributes dataset. Specifically, we use the VisA dataset (Silberer et al., 2013); (2) The Se-

Concept category	# of assigned words
animals	138
food	59
home	49
clothing	42
artefacts	38
vehicles	37
tools	27
structures	26
weapons	23
instruments	19
appliances	18
container	12
device	8
plants	7
material	4
toys	3

Table 1: VisA concepts by the categories

mantic Task Layer then fine-tunes the pretrained attribute vectors through supervised lexical entailment tasks (Section 2). Through the experiments (Section 4), we empirically confirmed that the resulting fine-tuned attribute vectors contributed to the improvement in semantic/visual similarity/relatedness tasks (Section 5). We also observed that the gold attribute annotations were recovered in reasonably good accuracies, and potentially good attributes were additionally discovered. These results may successfully address the issues with the existing work (Section 6), and contribute to the development of an automated method to acquire/refine semantic attributes that were originally captured by a human-involved annotation process.

VisA: The present study uses the VisA (Feris et al., 2017)[1] dataset as the source of human-annotated data instances, as well as the inventory of attributes. It is a dataset of images and their visual attributes for the nouns contained in McRae's feature norm dataset (McRae et al., 2005). Note however that we never use the collected images. These visual attributes were given to 510 noun concepts, rather than images, using 721 attribute types[2]. Each of them turns out to be associated with a component of an attribute vector. The 510 noun concepts are classified into 16 coarse categories, whereas the 721 attributes are grouped into 10 attribute groups as summarized in Table 1 and Table 2, respectively. As expected, the attributes exhibit a long-tailed distribution. Table 3 counts the occurrences of major attributes (frequency > 100) with the corresponding attribute groups, showing that the anatomy and color_patterns are indeed frequently used attribute groups. One of the merits of the VisA dataset, compared to McRae's feature norms, is that special care was taken to supply appropriate attributes that were missing in McRae's feature norms, although the completion may be imperfect.

[1] http://homepages.inf.ed.ac.uk/s1151656/resources.html

[2] Although this number is different from 636, publicized by (Silberer et al., 2013), it is the actual number calculated from the publicized dataset.

Attribute group	Used frequency
anatomy	1941
colour_patterns	1557
parts	1514
texture material	809
shape_size	738
behaviour	441
diet	358
botany	254
inbeh	69
structure	42

Table 2: VisA attributes by the groups

Attribute	Group	Used frequency
is_black	colour_patterns	238
is_brown	colour_patterns	224
is_white	colour_patterns	196
different_colours	colour_patterns	160
beh_-_eats	behaviour	138
made_of_metal	texture_material	130
has_eyes	anatomy	127
has_head	anatomy	121
beh_-_drinks_water	diet	110
has_tail	anatomy	108
has_tongue	anatomy	106
has_ears	anatomy	102

Table 3: VisA major attributes by the groups

2. Lexical Entailment Tasks for Fine-tuning

Lexical entailment is a semantic relation that holds between words, or more precisely, between concepts that the words in question denote. Formally, the entailment relation holds between concept x and y, if one of them can be inferred from the other. For example, a "dog" entails the existence of a "mammal."

Lexical entailment is clearly an *asymmetric* semantic relation; hence, the decision of directionality is a major matter of concern. In fact, many studies (Geffet and Dagan, 2005; Kotlerman et al., 2010; Gawron, 2014) have dealt with this issue by considering *distributional inclusion hypothesis*, which assumes that the set of properties denoted by a hyponym forms a subset of the hypernym properties. The rationale behind the present study is in line with this approach. That is, we expect that many of the attributes associated with a hyponym form a subset of the hypernym's attribute set. In fact, we verified this expectation with the VisA dataset (Silberer et al., 2013), where more than 70% of the attributes of hyponyms were inherited from their direct hypernyms. It would thus be possible to employ lexical entailment directionality detection task as a training task for predicting an adequate set of semantic attributes. In particular, if we are given an initial set of attribute vectors for the vocabulary of concern, such a directionality task can be utilized as a task for fine-tuning the attribute vectors.

3. Proposal: Word2Attr

Figure 1 presents a schematic view of the proposed Word2Attr network architecture, where the basic building blocks are two Attribute layers for pretraining and one Se-

mantic task layer for fine-tuning. As the two Attribute layers share their parameters, the whole network architecture forms a Siamese-style neural network (Koch et al., 2015), which has been applied to pairwise tasks, such as the prediction of similarity between two linguistic objects (He et al., 2015; Mueller, 2016).

3.1. Attribute Layer

The role of the Attribute layer is to initially map an input word embedding vector to its corresponding attribute vector. These layers are implemented by a one-hidden layer MLP. The mean squared error (MSE) between a predicted attribute vector $\hat{\boldsymbol{a}}$ and the gold attribute vector $\boldsymbol{a}$ is employed as the loss function. Note that the dimensionality of $\hat{\boldsymbol{a}}$ equals to the number of attribute types, and each component of a gold attribute vector is binary (0 or 1), indicating whether the corresponding attribute (e.g. `is_red`) is present or not.

$$\boldsymbol{h} = \text{Leaky\_ReLU}(W_a^{(1)}\boldsymbol{x} + b_a^{(1)}) \quad (1)$$

$$\hat{\boldsymbol{a}} = \text{sigmoid}(W_a^{(2)}\boldsymbol{h} + b_a^{(2)}) \quad (2)$$

This supervised task can be considered as an independent learning task in its own, but it is considered as a pretraining step in the whole learning process. That is, the learned parameters ($\boldsymbol{W}$ and $\boldsymbol{b}$) are used to initialize the corresponding parameters in the lexical entailment fine-tuning tasks.

3.2. Semantic Task Layer

As discussed, one of the reasons why we introduce a lexical entailment task as the fine-tuning task is the expectation that semantic attributes (or more precisely, the relationship between the sets of attributes) play a significant role in deciding the entailment directionality. Another reason to introduce a pairwise task, such as lexical entailment, is that it could partly reduce the issue of small-sized data.

The whole network architecture of Word2Attr, depicted in Figure 1, is inspired by the Supervised Directional Similarity Network proposed in a study (Rei et al., 2018), which deals with a graded lexical entailment task. Instead of learning to predict the degree of the entailment relation, we concentrate on training the network with the entailment directionality tasks. That is, given a pair of words (w_1, w_2), the network simply tries to assign a label, either of `hyper` or `hypo`.

Inputs to the Semantic task layer are vectors of the paired words. Each $\boldsymbol{x_i}$ is constructed by the weighted-concatenation of the word embedding vector $\boldsymbol{w_i}$ and the pretrained attribute vector $\hat{\boldsymbol{a}}_{\boldsymbol{i}}$ as follows:

$$\boldsymbol{x_i} = \rho_w \boldsymbol{w_i} \oplus \rho_a \hat{\boldsymbol{a}}_{\boldsymbol{i}} \quad (3)$$

Although ρ_w and ρ_a dictate weights for their corresponding vectors, we use them as an indicator for the inclusion of their vectors. More precisely, we compare three cases with the configurations: $(\rho_w, \rho_a) = \{(1,0),(0,1),(1,1)\}$.

To capture feature interactions between two words, their vectors are cross-multiplied by the following gating layers:

$$\boldsymbol{g_i} = \text{sigmoid}(W_{s_i}^{(1)}\boldsymbol{x_i} + b_{s_i}^{(1)}) \quad (4)$$

$$\tilde{\boldsymbol{x}}_{\boldsymbol{1}} = \boldsymbol{x_1} \circ \boldsymbol{g_2} \quad (5)$$

$$\tilde{\boldsymbol{x}}_{\boldsymbol{2}} = \boldsymbol{x_2} \circ \boldsymbol{g_1} \quad (6)$$

Here, $\circ$ denotes an element-wise multiplication operation.

The next layer is a mapping layer that accomplishes mapping into the space, which is more appropriate for directionality tasks. Each $W_{s_i}^{(2)}$ is expected to capture the asymmetry that is inherent to entailment relations.

$$\boldsymbol{m_i} = \tanh(W_{s_i}^{(2)}\boldsymbol{x_i} + b_{s_i}^{(2)}) \quad (7)$$

The resulting vectors are combined by element-wise multiplication, and subsequently fed into the MLP that is responsible for the final classification.

$$\boldsymbol{d} = \boldsymbol{m_1} \circ \boldsymbol{m_2} \quad (8)$$

$$\boldsymbol{h} = \tanh(W_s^{(3)}\boldsymbol{d} + b_s^{(3)}) \quad (9)$$

$$\hat{\boldsymbol{y}} = W_s^{(4)}\boldsymbol{h} + b_s^{(4)} \quad (10)$$

The overall loss function (Eq. 11) jointly considers the attribute vectors MSE loss (Eq. 12) and the directionality classification loss (Eq. 13) implemented by softmax cross entropy (indicated smxe in Eq. 13).

$$\mathcal{L} = \alpha\mathcal{L}_{attr} + \beta\mathcal{L}_{dir} \quad (11)$$

$$\mathcal{L}_{attr} = \frac{1}{N}\sum^{N}(\hat{\boldsymbol{a}}_{\boldsymbol{1}} - \boldsymbol{a_1})^2 + \frac{1}{N}\sum^{N}(\hat{\boldsymbol{a}}_{\boldsymbol{2}} - \boldsymbol{a_2})^2 \quad (12)$$

$$\mathcal{L}_{dir} = \frac{1}{N}\sum^{N}\text{smxe}(\hat{\boldsymbol{y}}, \boldsymbol{y_{\text{type}}}) \quad (13)$$

The impacts of these losses are adjusted by the weight parameters, α and β. Note that the attribute loss is set to zero if the corresponding gold attributes are not provided in the training data. In these cases, the network simply relies on the lexical entailment directionality signal.

4. Experimental Setup

This section describes the experimental settings in two supervised learning tasks, for pretraining and fine-tuning, and target similarity/relatedness tasks.

4.1. Pretraining of the Attribute Layer

We pretrained the Attribute layer using the following setup.

Word embedding vectors: We used the one million fastText (Bojanowski et al., 2016) pretrained word vectors[3] trained on Wikipedia 2017, UMBC webbase corpus, and statmt.org news datasets. The 300-dimensional vectors were L2-normalized.

Optimizer and activation function: We used Adam to optimize the learnable parameters. The parameter α of the leaky ReLU was set to 0.2.

Layer size: The dimensionality of the hidden layer was 512.

[3] `https://s3-us-west-1.amazonaws.com/fasttext-vectors/wiki-news-300d-1M.vec.zip`

Dataset: To accomplish the supervised learning task, we utilized the VisA dataset (Silberer et al., 2013). We discarded four nouns with multiple senses. We split the remaining 506 noun instances into training and test sets with 404 and 102 nouns, respectively. For polysemous words, we adopted the McRae's first senses.

4.2. Fine-tuning by Lexical Entailment Directionality Task

In addition to the standard directionality task (`dir`; binary classification), which is to predict the directionality (hypernym or hyponym) of a given word pair, we considered an extended directionality task (`ext`; ternary classification), which is to classify the entailment type of a given word pair as one among `hyper`, `hypo`, and `cohyp`. We state that the inclusion of co-hyponymy relation (`cohyp`) is reasonable, as this relation tends to be confusing with both hyponymy and hypernymy relations, thereby contributing towards a better training. We trained and tested the Semantic task layer using the following setup:

Optimizer: Adam was also used in this step.

Layer size: 512 for $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$; 256 for $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$; 64 for $\boldsymbol{h}$.

Weight parameters in the loss function: The weight parameters, α and β, in the loss function (Eq. 11) were set to 1.0 and 0.8, respectively.

Validation: We employed the F-1 measure in classification as the stopping criteria for the validation step.

Datasets: To accomplish the supervised learning task, we utilized **HyperLex** (Vulić et al., 2017) and **BLESS** (Baroni and Lenci, 2011), which are summarized below. We split the data instances into a training set (70 %), validation set (5 %), and test set (25 %), respectively.

HyperLex: The present study uses HyperLex (Vulić et al., 2017)[4] as the source of annotated lexical entailment data. This dataset collects 2,616 concept pairs that are human-annotated with the direction and the degree of entailment. We particularly focus on the directionality, as asymmetry is the central property of lexical entailment. In this study, we considered only noun-noun pairs, which amount to 2,163 pairs of instances. This dataset maintains seven semantic relation types: `hyp-N` (hypernym), `r-hyp-N` (hyponym), `cohyp` (co-hypernym), `mero` (meronym), `syn` (synonym), `ant` (antonym), and `no-rel` (no relations). Table 4 counts the number of pairs in each of the considered semantic relation types, and displays the corresponding examples. Note that Rand (random) and Lex (lexical) respectively indicate data split conditions.

Rel. type	Rand	Lex	Example
hyper	1,004	609	penguin - bird
hypo	226	127	reptile - alligator
cohyp	242	123	asia - africa
no-rel	160	95	enemy -crocodile
Total	1,632	954	

Table 4: Breakdown of the noun-noun portion of HyperLex

BLESS: This dataset[5] provides 14,400 tetrads of (target_concept, relatum, semantic relation type, topical domain type). It maintains six semantic relation types: `coord`, `hyper`, `mero`, `attri`, `event`, and `random`. Among these, `coord` is equivalent to HyperLex's `cohyp`, and `random` corresponds to `no-rel`. In this study, we extracted noun-noun pairs whose semantic relation type was either `hyper` or `coord`. Then, we artificially added reversed `hyper` word pairs as the equivalents of `hypo` pairs in HyperLex. Totally, we used 12,743 noun-noun pairs. Table 5 summarizes the noun-noun portion of the BLESS dataset. Note that we also included the `no-rel` (HyperLex) and `random` (BLESS) instances during the training.

Rel. type	# of pairs	Example
hyper	1,337	alligator - animal
hypo	1,337	animal - alligator
coord	3,565	alligator - lizard
random	6,702	alligator - handgun
Total	12,743	

Table 5: Breakdown of the noun-noun portion of BLESS

4.3. Evaluation by Semantic/Visual Similarity/Relatedness Tasks

To assess the efficacy of fine-tuning with the lexical entailment tasks, we applied the resulting attribute vectors to the target evaluation tasks, which were semantic/visual similarity/relatedness tasks. We evaluated the degree of similarity/relatedness by calculating the cosine of the two vectors.

Evaluation metrics: The fitness of the predictions against the gold annotations were measured by the Spearman's rank correlation coefficient.

Datasets: We utilized three datasets: **SemSim/VisSem** (Silberer and Lapata, 2014), **MEN** (Bruni et al., 2014), and **SimLex999** (Hill et al., 2015), which are described below.

SemSim/VisSim: This dataset, obtained from the same website as the VisA dataset, collects 7,576 word pairs, each of which is annotated using not only semantic similarities (SemSim), but also visual similarities (VisSim), so that users can compare the performances of their model in predicting different types of similarities. As the vocabulary of this dataset originates from McRae's feature norms, its coverage against the VisA dataset is very high.

[4] `http://people.ds.cam.ac.uk/iv250/hyperlex.html`

[5] `https://sites.google.com/site/geometricalmodels/shared-evaluation`

MEN: This dataset[6] includes 3,000 word pairs created from 751 distinct words. Each pair in the dataset is given a semantic relatedness score in the range of [0, 1]. It contains highly semantically related pairs (e.g., *beach/sand* rated as 0.96) as well as low-scored pairs (e.g., *bakery/zebra* rated as 0). Each word in the dataset is assigned a part of speech (POS) tag: verb, adjective, or noun. We used a subset of 2,005 noun-noun pairs (645 words). Its coverage against the VisA dataset was very low: 139/645 (21.6 %) in words and 101/2,005 (5 %) in pairs.

SimLex999: This dataset provides word similarity (rather than relatedness or association) judgments for 999 word pairs. Note that the parts of speech of compared words are always the same (noun, adjective, and verb). We used a subset of 666 noun-noun pairs out of 751 distinct nouns. Its coverage against the VisA dataset was very low: 90/751 (12 %) in words and 43/666 (6 %) in pairs.

5. Results and Discussion

This section first describes the experimental results of the lexical entailment directionality tasks. These are not the target tasks to assess the efficacy of predicted attributes. Nevertheless, it is necessary to confirm that reasonably good results can be achieved by these tasks. We then discuss the results of the target semantic/visual similarity/relatedness tasks. Finally, we discuss how the proposed method could or could not recover the human-annotated gold attributes in VisA by presenting relevant examples.

5.1. Lexical Entailment Directionality

Table 6 summarizes the results of the lexical entailment directionality tasks with the HyperLex dataset. Notice that the results with the BLESS dataset are not shown. Due to the nature of the BLESS dataset, we could not attain the lexical split, which lead to nearly perfect (close to 100%) results. These results may attribute to the issue of *prototypical* word (Levy et al., 2015).

As described earlier, there are two tasks: directionality task (`dir`; binary classification), and extended directionality task (`ext`; ternary classification, including `cohyp`). We compared the precision/recall/F1 (P/R/F1) scores of the three input vectors (fText, attr, and fText+attr) in two HyperLex data split settings (random and lexical).

- Input vectors:

  fText: Only fastText (Bojanowski et al., 2016) word embedding vectors were utilized.

  attr: Only pretrained attribute vectors were employed.

  fText+attr: Two vectors were concatenated as formulated in Eq. 3.

- Data split: The following standard splits are provided in the HyperLex dataset:

  Random split: The train/validation/test subsets are selected by random sampling, and no word pair overlaps are allowed among these subsets.

  Lexical split: The train and test subsets are rigorously constructed to ensure "zero lexical overlap" by discarding all "cross-set" training-test concept pairs, insisting that this data split is more difficult than the random split.

Table 6 demonstrates the efficacy of attribute vectors in the given task settings. Consistent improvements in the P/R/F1 scores were observed with the attr and fText+attr input vectors when compared to the fText baseline. These results suggest that even the error-prone predicted attribute vectors could play a role in capturing useful information for deciding entailment directionality.

Notable results were found with the lexical split setting. The score degradation in the fText baseline was highly prominent when compared to the results with other inputs, suggesting that the attribute vectors play a vital role in capturing some semantic aspects of unseen words. These promising results corroborate that the lexical entailment tasks can be effectively used for the fine-tuning of attribute vectors.

5.2. Visual/Semantic Similarity/Relatedness

Table 7 compares Spearman's correlation coefficients acquired from several experimental settings in the visual/semantic similarity/relatedness prediction tasks.

- Training datasets used in the lexical entailment tasks: **BLESS**, **HyperLex (random split)**, and **HyperLex (lexical split)**.

- Input vectors: fText, VisA binary attr, pretrain attr (baseline), and fText+attr.

  VisA binary attr: Gold annotations in the VisA dataset were directly used as binary attribute vectors.

  Word2Attr pretrain (baseline): Attribute vectors only pretrained by the Attribute layer of the Word2Attr architecture.

In Table 7, the percentages (100, 5, and 6%) indicate the coverage of the word pairs in the corresponding dataset. Unseen words were used in the settings denoted by 100% in MEN and SimLex, providing more difficult cases.

As displayed in the table, the proposed method performed better in the SemSim and VisSim datasets than almost other methods, including the previous work (Făgărăşan et al., 2015; Bulat et al., 2016). These results suggest that fine-tuning with the lexical entailment task was indeed effective. Interestingly, the predicted attributes (originated from visual attributes annotated in ViSA) contributed to the performance improvement in VisSim and SemSim, outperforming not only the existing results but also that with fastText. However, significant differences were not observed between the `dir` and `ext` tasks.

The results with MEN and SimLex, on the other hand, insist that the fastText embeddings are remarkably robust. An obvious reason for this is that they are induced from a huge corpora. Another reason can be attributed to the nature of the datasets, particularly the *concreteness* of the included

[6] `https://staff.fnwi.uva.nl/e.bruni/MEN`

		HyperLex (random)						HyperLex (lexical)					
		dir			ext			dir			ext		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
fText	total	0.89	0.86	0.87	0.79	0.79	**0.79**	0.66	0.81	0.73	0.44	0.66	0.53
	hyper	0.94	0.93	0.93	0.84	0.92	0.88	0.81	1.0	0.9	0.66	1.0	0.8
	hypo	0.67	0.56	0.61	0.81	0.67	0.73	0.0	0.0	0.0	0.0	0.0	0.0
	cohyp	-	-	-	0.55	0.38	0.45	-	-	-	0.0	0.0	0.0
attr	total	0.92	0.87	0.88	0.82	0.77	**0.79**	0.85	0.78	**0.82**	0.68	0.65	0.65
	hyper	0.93	0.93	0.93	0.87	0.87	0.87	0.91	0.86	0.88	0.78	0.78	0.77
	hypo	0.87	0.57	0.68	0.87	0.6	0.7	0.62	0.51	0.56	0.5	0.38	0.42
	cohyp	-	-	-	0.54	0.49	0.51	-	-	-	0.44	0.4	0.4
fText+attr	total	0.92	0.86	**0.89**	0.72	0.75	0.72	0.85	0.79	0.81	0.68	0.66	**0.66**
	hyper	0.95	0.9	0.93	0.81	0.92	0.86	0.9	0.86	0.88	0.78	0.81	0.79
	hypo	0.78	0.67	0.72	0.69	0.48	0.56	0.64	0.46	0.54	0.51	0.38	0.43
	cohyp	-	-	-	0.39	0.26	0.31	-	-	-	0.46	0.36	0.38

Table 6: Results of lexical entailment directionality task (Precision / Recall / F1-score)

Train. Dataset	Input	Train Task	SemSim 100%	VisSim 100%	MEN 100%	MEN 5%	SimLex 100%	SimLex 6%
		fastText	0.66	0.56	**0.82**	**0.80**	**0.49**	0.47
		VisA binary attr	0.69	0.59	NA	0.64	NA	0.49
		Word2Attr pre-train (baseline)	0.73	0.63	0.62	**0.72**	0.38	**0.56**
BLESS	a	dir	0.75	0.64	0.67	0.7	0.42	0.55
		ext	0.76	0.64	0.65	0.67	0.4	0.55
	f+a	dir	0.76	0.65	**0.68**	0.71	**0.43**	**0.56**
		ext	**0.77**	0.65	0.66	0.69	0.41	**0.56**
HyperLex (random)	a	dir	0.75	**0.66**	**0.68**	**0.72**	0.4	0.54
		ext	0.74	0.65	0.67	0.71	0.39	0.54
	f+a	dir	0.75	**0.66**	**0.68**	0.71	0.39	0.52
		ext	0.75	**0.66**	**0.68**	**0.72**	0.41	0.55
HyperLex (lexical)	a	dir	0.74	0.65	**0.68**	0.7	0.39	0.53
		ext	0.74	**0.66**	**0.68**	**0.72**	0.4	0.53
	f+a	dir	0.76	0.65	**0.68**	**0.72**	0.38	0.53
		ext	0.74	0.65	**0.68**	**0.72**	0.4	**0.56**
Făgărăşan et al. (Fagarasan2015) (reproduced)			0.75	0.61	**0.68**	0.71	0.4	0.45
Bulat et al. (Bulat2016) (reproduced)			0.74	0.61	**0.68**	0.69	0.42	0.46

Table 7: Results of semantic/visual similarity/relatedness prediction (Spearman's correlation)

concepts. In fact, the average concreteness scores (Brysbaert et al., 2014) for MEN and SimLex were 4.61 and 4.05, respectively, which are lower than that of VisSim, 4.83. These concreteness analyses suggest that the source of attributes should be largely extended to accommodate more abstract words.

5.3. Recovery of the Gold Attributes

Although the recovery of the attributes presented in the dataset is not our primary goal, it would be worth seeing how well the proposed method recovered them. By applying a threshold to the components of an attribute vector, we can recover a set of semantic attributes from the attribute vector, which can be directly compared with the gold attributes. The best results we obtained in the macro-averaged precision, recall, and F1 were: 0.90 / 0.84 / 0.87 for the training data, and 0.65 / 0.5 / 0.56, for the test data, respectively. Given that the annotations in the VisA dataset are largely incomplete and somewhat noisy (Făgărăşan et al., 2015), these results could be considered modestly good. To further investigate which attributes of which concepts are difficult to recover, we computed the F1 score differences between the training data and the test data. We can assume that the bigger difference may allude that the corresponding concept/attribute is difficult to generalize. Figure 2 presents the results in a heatmap, where the x-axis and the y-axis respectively classify attribute groups and concept categories. Remind that the concept-attribute combinations not in the VisA data sets are also included in the heatmap. We found that concept categories, such as device, instruments, artefacts, and appliances present more difficulty to recover their attributes. This trend may indicate that the concepts chiefly characterized by their functions and/or constructs are relatively difficult to generalize to predict. We also noticed that the attributes like shape size, texture material, and colour patterns were difficult to be recovered. This outcome could be attributed to the fact that many of the concrete objects can appear in multiple colors and in various texture, and the gold annotations might have reflected this diversity.

5.4. Examples of the Predicted Attributes

We observed that potentially reasonable and effective attributes were additionally discovered in the predicted attributes. Table 8 classifies the Word2Attr-predicted attributes for a selected set of words. For each word, the pre-

Word	Status	Attributes
alligator	extra	**has_jaws***, **has_warts***, **beh_-_eats_fish***, **is_yellow**, **beh_-_walks***, **beh_-_eats_plants***
	gold	**has_4_legs**, has_ears, **has_powerful_jaws**, beh_-_eats, has_tongue beh_-_swims, is_long, **has_neck**, is_green, beh_-_eats_small_animals has_head, has_mouth, has_feet, has_teeth, has_toes has_eyes, has_snout, has_nose, is_brown, beh_-_drinks_water
	lack	*has_tail*, is_black, is_grey, is_large, has_scales, has_claws, beh_-_crawls
axe	extra	made_of_metal*, **made_of_iron***, has_edge*, **made_of_wood***
	gold	has_handle, made_of_steel, is_flat, has_blade, is_silver is_brown, has_long_handle, has_wooden_handle, **has_head**
	lack	is_T_shaped, different_shapes, is_black, has_metal_head, is_L_shaped
celery	extra	**is_leafy***, **has_top**, **is_brown**, has_peel, **has_skin**, **has_pointed_end** is_cylindrical*, has_seeds, has_layers
	gold	is_long, has_stalks, is_green, has_leaves
	lack	has_stem
trumpet	extra	**is_silver***
	gold	is_gold, **is_shiny**, **has_mouthpiece**, **has_slide**, **has_tubing**
	lack	has_bell, has_brace, has_cylindrical_bore, made_of_brass has_ring, has_valves, has_buttons
lantern	extra	-
	gold	-
	lack	is_black, has_stand, made_of_metal, has_light, has_candle, is_brown, made_of_glass

Table 8: Examples of the predicted attributes (th=0.93).(Setups: HyperLex random / attr / dir)

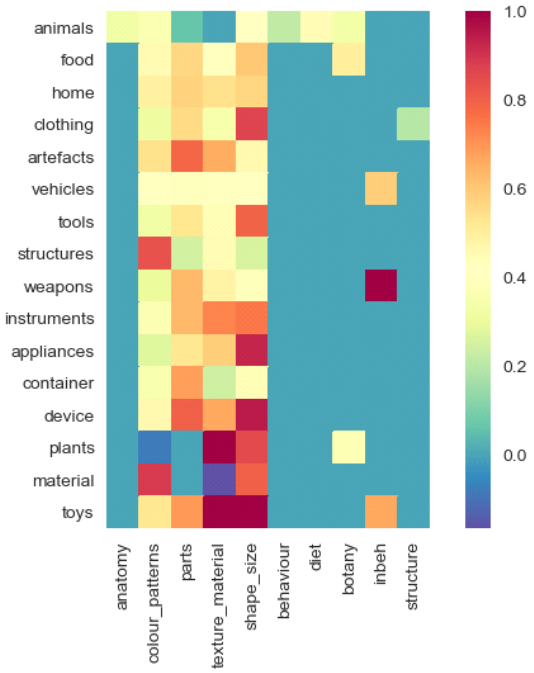

Figure 2: Train-Test F1-score differences.

dicted attributes are classified into one of the three groups: `extra`, `gold`, and `lack`.

- `extra`: Additionally predicted attributes that are not included in the original set of attributes are listed. Of these, the attributes judged relevant by the authors are marked with the asterisk. Besides, the attributes predicted only through the fine-tuning process are shown in bold.
- `gold`: The attributes given in the ViSA dataset and predicted by the Word2Attr model are shown. Note that the ones in the plain font are only recovered by the pre-training process.
- `lack`: The gold attributes that could not be recovered by the predictions are displayed. The italicized attributes (*has_tail* for alligator only) are the ones predicted by the pretrainig, but not by the fine-tuning.

The overall tendencies that can be drawn from the observation of actual data is two fold: (1) the fine-tuned vectors tends to predict additional attributes, where some of them are relevant, and (2) the pre-trained vectors could steadily recover the gold attributes. These results suggest that we would be able to better combine these different types of vectors.

6. Related Work

Utilization of semantic feature norms: McRae's semantic feature norm dataset is a seminal resource that facilitates a line of researches, highlighting the nonlinguistic aspects of meanings. One of the pioneering works (Silberer and Lapata, 2012) employed McRae's feature norms as a proxy for perceptual information. Subsequently, the work (Silberer and Lapata, 2014) integrated perceptual features, mainly acquired from images, with linguistic features. Most of them demonstrated the effectiveness of nonlinguistic semantic features in semantic similarity/relatedness tasks.

Automatic acquisition of semantic attributes: McRae's dataset suffers from its coverage (it only covers a small set of basic-level concepts) and completeness (often evidently relevant features are missing). To deal with these issues, several trials have been made (Făgărăşan et al., 2015; Bulat et al., 2016), including the present study. However, only a few studies utilized semantic attributes/features in downstream NLP applications. Among these, Lazaridou2016 (Lazaridou2016) were the first to apply semantic

attributes/features in differentiating the meaning of a concept from others. As a closely related but different approach is recently proposed in (Derby et al., 2019), where a mapping from the feature domain onto the existing word embedding space is learned.

Lexical entailment as an asymmetric semantic relation: Lexical entailment, which was employed as the fine-tuning task, can be discussed in the broader context of asymmetric semantic relationships (Kotlerman et al., 2010; Gawron, 2014). Most of the studies (Geffet and Dagan, 2005) assumed the feature inclusion hypothesis in capturing the relational directionality, and this idea was also inherited in our study. The present study may be the first attempt to incorporate the tasks of lexical entailment in learning to capture word attributes. Recently, the development of the HyperLex dataset (Vulić et al., 2017) invoked the work on the estimation of graded lexical entailment (Vulić and Mrkšić, 2017; Rei et al., 2018). Our work, however, has been centered on the directionality task, as it may be essential in refining existing attribute-based representations.

Contextualized word representations: As demonstrated by the impressive successes of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), the contextualized word representations might be better employed than the attribute-based representations in certain types of NLP tasks. Besides, these word representations may successfully capture attribute information not explicitly as in this study, but implicitly. The present study has not compared the attribute vectors with contextualized representations, as the so far employed evaluation tasks are context-independent similarity/relatedness tasks. We should, however, seek some application areas that could be appropriate for the comparison.

7. Conclusion

Given the potential applicability of attribute-based vector representation of words, an automatic method to develop and maintain a language resource of the type is highly demanded. We proposed Word2Attr, a neural network model that can predict attribute vectors even for an unseen word. We empirically demonstrated that the resulting attribute vectors contributed to the improvement in semantic/visual similarity/relatedness tasks, emphasizing that the fine-tuning process that employs the lexical entailment tasks could successfully capture some useful aspects of word concepts. These results presumably shed light on the development of an automated method to acquire/refine semantic attributes, and may contribute to the development and maintenance of attribute-based semantic resources.

The present results, on the other hand, highlighted possible lines of future work that include: (1) incorporating external knowledge presumably from a knowledge graph (Speer et al., 2017) to enrich background information for dealing with the functional aspects of a concrete object, as well as more abstract concepts, and (2) integrating with the features induced from non-textual sources, such as images (Kiela et al., 2016; Hasegawa et al., 2017), to capture the perceptual aspects of a concrete object.

As a more fundamental issue, we should develop a method to induce an appropriate set of attributes from various information sources and to organize them as a structured inventory of attributes. Such a method should consider both the hierarchical and disjunctive natures in some attributes. For example, the attribute `has_strong_jaws` should be organized as a sub-attribute of `has_jaws`, whereas `has_color` for apple can be exapanded into either of `is_red` or `is_green`.

Acknowledgements

The present work was partially supported by JSPS KAKENHI Grants number 17H01831.

8. Bibliographical References

Al-Halah, Z., Tapaswi, M., and Stiefelhagen, R. (2016). Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. *Proceedings of CVPR 2016*, pages 5975–5984.

Baroni, M. and Lenci, A. (2011). How we BLESSed distributional semantic evaluation. *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics (GEMS)*, pages 1–10.

Barsalou, L. W. (2008). Grounded cognition. *Annual review of psychology*, 59(August):617–645.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bruni, E., Gatica-perez, D., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(December):1–47.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–11.

Bulat, L., Kiela, D., and Clark, S. (2016). Vision and Feature Norms : Improving automatic feature norm learning through cross-modal maps. *Proceedings of NAACL-HLT 2016*, pages 579–588.

Derby, S., Miller, P., and Devereux, B. (2019). Feature2Vec: Distributional semantic modelling of human property knowledge. 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 5852–5858, Hong Kong, China, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, abs/1810.04805.

Rogerio Schmidt Feris, et al., editors. (2017). *Visual Attributes*. Springer.

Făgărăşan, L., Vecchi, E. M., and Clark, S. (2015). From distributional semantics to feature norms : grounding semantic models in human perceptual data. *Proceedings of IWCS 2015*, pages 52–57.

Gawron, J. M. (2014). Improving sparse word similarity models with asymmetric measures. *Proceedings of ACL 2014*, pages 296–301.

Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. *Proceedings of ACL 2005*, pages 107–114.

Hasegawa, M., Kobayashi, T., and Hayashi, Y. (2017). Incorporating visual features into word embeddings : A bimodal autoencoder-based approach. *Proceedings of IWCS 2017*.

He, H., Gimpel, K., and Lin, J. (2015). Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. *Proceedings of EMNLP 2015*, pages 1576–1586.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Kiela, D., Verő, A. L., and Clark, S. (2016). Comparing data sources and architectures for deep visual representation learning in semantics. *Proceedings of EMNLP 2016*, pages 447–456.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. *Proceedings of ICML 2015 Deep Learning Workshop*.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Krebs, A., Lenci, A., and Paperno, D. (2018). SemEval-2018 task 10: Capturing discriminative attributes. *Proceedings of SEMEVAL 2018*, pages 732–740, June.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2014). Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465.

Lazaridou, A., Pham, N. T., and Baroni, M. (2016). The red one!: On learning to refer to things based on their discriminative properties. *Proceedings of ACL 2016*.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015). Do Supervised Distributional Methods Really Learn Lexical Inference Relations? *Proceedings of NAACL-HLT 2015*, pages 970–976.

Li, K. C., Zhang, J., Zhang, J., and Huang, K. (2018). Discriminative learning of latent features for zero-shot recognition. *Proceedings of CVPR 2018*, pages 7463–7471.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of NIPS 2013*, 9:3111–3119.

Mueller, J. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. *Proceedings of AAAI 2016*, pages 2786–2792.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe : Global Vectors for Word Representation. *Proceedings of EMNLP 2014*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*.

Rei, M., Gerz, D., and Vulić, I. (2018). Scoring Lexical Entailment with a Supervised Directional Similarity Network. *Proceedings of ACL 2018*.

Silberer, C. and Lapata, M. (2012). Grounded models of semantic representation. *Proceedings of EMNLP-CoNLL 2012*, pages 1423–1433.

Silberer, C. and Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. *Proceedings of ACL 2014*, pages 721–732.

Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of Semantic Representation with Visual Attributes. *Proceedings of ACL 2013*, pages 572–582.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of AAAI 2017*.

Vulić, I. and Mrkšić, N. (2017). Specialising Word Vectors for Lexical Entailment. *Proceedings of NAACL-HLT 2017*, pages 1134–1145.

Vulić, I., Gerz, D., Kiela, D., Hill, F., and Korhonen, A. (2017). HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.