Minimally-Supervised Joint Learning of Event Volitionality and Subject Animacy Classification

Hirokazu Kiyomaru, Sadao Kurohashi

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{kiyomaru, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Volitionality and subject animacy are fundamental and closely related properties of events. Their classification, however, is challenging because it requires contextual text understanding and a huge amount of labeled data. This paper proposes a novel method that jointly learns volitionality and subject animacy at a low cost, heuristically labeling events in a raw corpus. Volitionality labels are assigned using a small lexicon of volitional and non-volitional adverbs such as deliberately and accidentally; subject animacy labels are assigned using a list of animate and inanimate nouns obtained from ontological knowledge. Since our labeling method assigns labels only to a biased set of events, a classifier is trained with regularization to take into account the property. This paper explores the following two approaches: bias reduction and adversarial representation learning. In bias reduction, the words used for labeling are regarded as bias that should not be overexploited to make predictions, and their estimated contribution towards predictions is penalized. In adversarial representation learning, the classifier is given unlabeled events as well and makes their latent representations closer to labeled events' ones in an adversarial manner while learning classification on labeled events. We conduct experiments with crowdsourced gold data in Japanese and English and show that our method effectively learns volitionality and subject animacy without manually labeled data.

1 Introduction

Volitionality is a fundamental property of events, which indicates whether an event represents an action that someone is volitionally involved in. For example, eating and writing are usually volitional; crying and getting injured are non-volitional. Event volitionality classification has been used for causal knowledge categorization (Lee and Jun 2008; Inui, Inui, and Matsumoto 2003; Abe, Inui, and Matsumoto 2008a,b) and has various potential applications such as conditional event prediction (Du et al. 2019), script induction (Chambers and Jurafsky 2008), and customer feedback analysis (Liu et al. 2017).

On the other hand, *animacy* is a fundamental property of nouns, which indicates whether the entity described by a noun is capable of human-like volition (Bowman and Chopra 2012). Animacy is closely related to volitionality because animacy is a necessary condition for volitionality. This suggests a synergistic effect of joint learning.

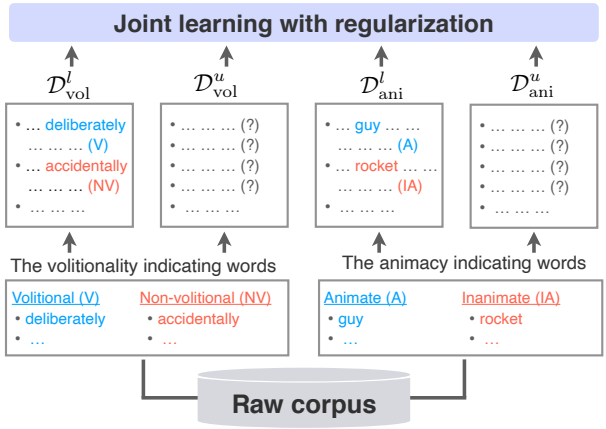

Figure 1: Overview of our method. We construct labeled/unlabeled datasets for volitionality/subject animacy classification by heuristically labeling events in a raw corpus using the volitionality/animacy indicating words. Our model jointly learns volitionality and subject animacy on them with regularization.

The challenge of identifying volitionality and animacy lies in limited language resources and contextual dependence. The volitionality of an event is largely decided by its predicate. However, existing language resources such as ConceptNet (Speer, Chin, and Havasi 2017) do not provide an exhaustive list of volitional predicates.

Even with a rich language resource, however, due to its context-dependent nature, volitionality cannot be entirely identified. Let us consider the following Japanese examples.

(1) a. *shawa-o* *abiru* (V)
shower-ACC[1] take
b. *hinan-o* *abiru* (NV)
criticism-ACC get

Examples (1-a) and (1-b) have the same predicate "*abiru* (take/get)," but the former is volitional, while the latter is

[1] ACC is the accusative case marker.

non-volitional.[2]

Similarly, example (2-a) is non-volitional, but example (2-b) is volitional because of the adverb "*fukaku* (deeply)."

(2) a. *iki-o* *suru* (NV)
breath-ACC take
b. *fukaku* *iki-o* *suru* (V)
deeply breath-ACC take

Coupled with the unbounded combinatorial nature of language, such contextual dependence entails the demand for learning from a huge amount of labeled data.

As for animacy, although there exist some available language resources listing animate/inanimate nouns, they are far from exhaustive. Besides, identifying animacy also requires contextual text understanding. For example, although examples (3-a) and (3-b) have the same subject "*shirobai* (white motorcycle)," the former describes an inanimate entity, a motorcycle, while the latter describes an animate entity, a police officer, as metonymy.[3]

(3) a. *shirobai-ga* *tometearu* (IA)
white motorcycle-NOM[4] be parked
b. *shirobai-ga* *oikaketekuru* (A)
white motorcycle-NOM chase

This paper proposes a minimally supervised method to jointly learn volitionality and subject animacy (Figure 1). We first assign labels to events in a raw corpus in a heuristic manner. Volitionality labels are assigned using a small lexicon of volitional and non-volitional adverbs, collectively called the *volitionality indicating words*. For example, example (4) is regarded as volitional because the volitional adverb "*aete* (deliberately)" modifies the predicate.

(4) *aete* *shinjitsu-o* *hanasu* (V)
deliberately truth-ACC tell

Example (5) is regarded as non-volitional because the non-volitional adverb "*ukkari* (accidentally)" modifies the predicate.

(5) *ukkari* *keitai-o* *otosu* (NV)
accidentally mobile-ACC drop

Subject animacy labels are assigned using a list of animate/inanimate nouns, collectively called the *animacy indicating words*, obtained from ontological knowledge. By using this labeling method, a large number of labeled events can be collected at a low cost.

As examples (4) and (5) suggest, we can consider that volitionality is preserved after removing the volitionality indicating words in most cases. The same can be said for subject animacy.

However, this is not always true. For example, example (6-a) is volitional, but example (6-b) is non-volitional. Here, the volitionality indicating word "*aete* (deliberately)" plays an essential role.

(6) a. *aete* *kokeru* (V)
deliberately tumble
b. *kokeru* (NV)
tumble

Such cases also exist in subject animacy classification. While the subject of example (7-a) "*shogeki* (impact)" is inanimate, the omitted subject of example (7-b) is normally assumed to be animate.

(7) a. *shogeki-ga* *hashiru* (IA)
impact-NOM run
b. *hashiru* (A)
run

In this paper, we learn a classifier with regularization to take into account this property. Specifically, we explore the following two approaches: bias reduction and adversarial representation learning. In bias reduction, the volitionality/animacy indicating words are regarded as bias and should not be over-exploited to make predictions. During training, the classifier learns to reduce the contribution of the volitionality/animacy indicating words towards predictions (Kennedy et al. 2020; Jin et al. 2020). In adversarial representation learning, the classifier is given unlabeled events as well and learns to make their latent representations closer to labeled events' ones using an adversarial learning framework while learning classification on labeled events (Ganin and Lempitsky 2015; Ganin et al. 2016).

We conduct experiments with crowdsourced gold data in Japanese and English and verify the effectiveness of the proposed method to learn volitionality and subject animacy without manually labeled data.

Our code and crowdsourced gold data are available at https://github.com/hkiyomaru/volcls.

2 Related Work

Our work mainly builds on event volitionality classification, bias reduction, and unsupervised domain adaptation.

2.1 Event Volitionality Classification

Previous work on event volitionality classification can be categorized into a targeted setting and a non-targeted setting. In the targeted setting, a model is given the predicate and its argument of an event and predicts whether the argument is volitionally involved in the action or state the predicate represents. This setting has been tackled as a sub-task of semantic proto-role labeling (Reisinger et al. 2015; White et al. 2016; Teichert et al. 2017).

In the non-targeted setting, which we tackle in this paper, a model is given an event and predicts whether the subject is volitionally involved in the event. To this end, Abe, Inui, and Matsumoto (2008a) and Abe, Inui, and Matsumoto (2008b) manually built a lexicon of verbs with volitionality labels and classified event volitionality by looking it up. This

[2]We use "V" and "NV" to indicate that an event is volitional and non-volitional from the viewpoint of the subject, respectively.

[3]We use "A" and "IA" to indicate that the subject of an event is animate and inanimate, respectively.

[4]NOM is the nominative case marker.

method is constrained by its inability to take context into account; as examples (1-a) and (1-b) suggest, volitionality depends on context.

Inui, Inui, and Matsumoto (2003) proposed a data-driven approach; they learned an SVM with hand-crafted linguistic features of events on a small amount of manually labeled data. However, the non-compositionality of event volitionality prevents us from learning from a small dataset. We use a massive amount of heuristically labeled events to learn a wide range of language phenomena and world knowledge related to volitionality.

2.2 Bias Reduction

Bias reduction is a technique to prevent a model from exploiting a specific bias to make predictions. While bias reduction has been actively studied in the field of fairness in machine learning (Bolukbasi et al. 2016; Zhao et al. 2017, 2019; Kennedy et al. 2020), we use this technique to prevent our model from over-exploiting the volitionality/animacy indicating words. Specifically, we employ two bias reduction methods proposed in Kennedy et al. (2020): word removal and explanation regularization based on sampling and occlusion (Jin et al. 2020). These methods were originally proposed to learn a hate speech classifier robust to group identifiers such as "gay." The details of these methods are deferred to Section 4.3.

2.3 Unsupervised Domain Adaptation

It is reasonable to employ semi-supervised learning techniques to solve our problem because our training data includes both labeled and unlabeled events. In the context of semi-supervised learning, given that our primary focus is on classifying unlabeled events to which our heuristics cannot assign labels, it is natural to view our problem as an unsupervised domain adaptation problem (Ramponi and Plank 2020).

Unsupervised domain adaptation is a technique to learn a model that better performs on a target domain, using labeled data from a source domain and unlabeled data from the target domain. We employ this technique regarding labeled events and unlabeled events as source domain data and target domain data, respectively. Specifically, we adopt adversarial domain adaptation (ADA) that has been used successfully in NLP tasks, including text classification such as sentiment analysis (Ganin et al. 2016; Ganin and Lempitsky 2015; Shah et al. 2018; Shen et al. 2018). In ADA, a model learns a latent feature space to reduce the discrepancy between the source and target distributions while learning a task using the source domain data, using an adversarial learning framework. The detail is deferred to Section 4.3.

3 Problem Setting

This section describes the representation, scope, and annotation of events we target in the present paper.

3.1 Representation

We represent an event as a clause, that is, a text that contains one main predicate. Compared to structured representations such as predicate-argument structures (Gildea and Jurafsky 2000), clauses can more flexibly represent the meaning of events. Besides, by representing events by clauses, we can obtain powerful event representations using strong pretrained text encoders like BERT (Devlin et al. 2019).

3.2 Scope

This paper focuses on events whose volitionality cannot be identified by simple linguistic features: POS tags and voice. We use POS tags to filter out events whose predicates are either an adjective or copula because they always represent a state and thus never represent a volitional action, as shown in examples (8) and (9).

(8) *sora-ga* *kireida* (NV)
sky-NOM be beautiful

(9) *kare-wa* *gakuseida* (NV)
he-NOM be student

As for voice, we filter out events in the passive or potential voice because they are not volitional from the viewpoint of their subjects, as shown in examples (10) and (11).

(10) *sensei-ni* *shikarareru* (NV)
teacher-DAT[5] be scolded

(11) *watashi-wa* *hashireru* (NV)
I-NOM can run

Besides, we filter out events with modality, linguistic expressions representing the writer's opinions or attitudes towards an event. Example (12) contains the modality of CERTAINTY expressed by "*hazuda* (should)."

(12) *kare-wa* *kuru* *hazuda*
he-NOM come should

Because our focus is on recognizing the volitionality of an event itself, we exclude such an event from the scope.

3.3 Annotation

An event is given volitionality and subject animacy labels.

Volitionality An event is considered *volitional* if the subject is volitionally involved in the event. Otherwise, it is considered *non-volitional*.

Subject Animacy The subject of an event is considered *animate* if the entity described by it can take volitional actions. Otherwise, it is considered *inanimate*. Since we consider a model that is given an event and predicts its subject animacy, we tie an animacy label to an event rather than the subject.

4 Proposed Method

Our goal is to train a model that is given an event x and predicts its volitionality y_{vol} and subject animacy y_{ani}. Both y_{vol} and y_{ani} take the value of 1 if positive (volitional/animate) and 0 if negative (non-volitional/inanimate). First, labeled events are collected from a raw corpus with our heuristic

[5]DAT is the dative case marker.

labeling method. Then, considering the property of the labeled events discussed in Section 1, our model jointly learns volitionality and subject animacy with regularization.

4.1 Constructing Training Dataset

We construct four types of datasets: events with volitionality labels $\mathcal{D}^l_{\text{vol}}$, events without volitionality labels $\mathcal{D}^u_{\text{vol}}$, events with subject animacy labels $\mathcal{D}^l_{\text{ani}}$, and events without subject animacy labels $\mathcal{D}^u_{\text{ani}}$.

First, events that satisfy the conditions described in Section 3.2 are extracted from a raw corpus, using an off-the-shelf syntactic dependency parser and POS tagger. Each of the events is then given its volitionality and subject animacy labels by our heuristic labeling method. According to the given label, the event is added to the corresponding dataset.

To assign the volitionality label, we prepare a small lexicon of volitional and non-volitional adverbs. If an adverb in the lexicon modifies the predicate of the event, the event is given the corresponding label and added to $\mathcal{D}^l_{\text{vol}}$. Otherwise, the event is added to $\mathcal{D}^u_{\text{vol}}$ without being given a label.

To assign the subject animacy label, we first find the subject of the event using a semantic dependency parser. If the subject is found, its animacy is then examined by looking up the animacy indicating words obtained from ontological knowledge and using the result of named entity recognition. If the animacy is identified, the event is associated with the corresponding label and pushed into $\mathcal{D}^l_{\text{ani}}$. If the subject is not found — which is not rare in pro-drop languages, including Japanese — or its animacy is not identified, the event is added to $\mathcal{D}^u_{\text{ani}}$ without being given a label.

4.2 Model

Our model consists of the following three neural networks: a text encoder E, a volitionality classifier C_{vol}, and a subject animacy classifier C_{ani}. The text encoder transforms an event x into a distributed representation. The volitionality classifier is given the representation and predicts the probability of x being volitional. Likewise, the subject animacy classifier predicts the probability that the subject of x is animate.

4.3 Training with Regularization

Our model jointly learns volitionality classification and subject animacy classification with regularization. As their training is done in a unified manner, we introduce placeholders for convenience. We refer to a labeled dataset as $\mathcal{D}^l$, an unlabeled dataset as $\mathcal{D}^u$, the label assigned to events in $\mathcal{D}^l$ as y, and the classifier to predict y as C. When learning volitionality, these placeholders are accompanied by the suffix "vol"; as for subject animacy, they are accompanied by the suffix "ani."

Our model learns classification using the labeled dataset. Formally, the objective is written as follows:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{(x,y)\sim\mathcal{D}^l}\text{BCE}(y, C(E(x))), \tag{1}$$

where BCE is binary cross-entropy.

We explore the following regularization methods.

Word Removal (WR) WR is a bias reduction method that decreases reliance on a word to make predictions by removing the word from training data. We use this method to reduce reliance on the volitionality/animacy indicating words. The objective is written as follows:

$$\mathcal{L}_{\text{WR}} = \mathbb{E}_{(x,y)\sim\mathcal{D}^l}\text{BCE}(y, C(E(x\backslash w))), \tag{2}$$

where w is the volitionality/animacy indicating word in x and $x\backslash w$ is x from which w is removed.

Explanation regularization by sampling and occlusion (SOC) SOC is a bias reduction method that penalizes the context-independent contribution of a word towards predictions (Kennedy et al. 2020). In order to estimate a context-independent contribution, SOC calculates the difference of model output after masking out the word, marginalized over all the possible context of the word. We use this method to reduce reliance on the volitionality/animacy indicating words. Formally, the objective is written as follows:

$$\mathcal{L}_{\text{SOC}} = \mathbb{E}_{x\sim\mathcal{D}^l}[\phi(x)]^2, \tag{3}$$

$$\phi(x) = \frac{1}{|\mathcal{S}|}\sum_{x'\in\mathcal{S}}[C(E(x')) - C(E(x'\backslash w))]^2, \tag{4}$$

where w is the volitionality/animacy indicating word in x, $\mathcal{S}$ is a set of events created by sampling the context of w according to a pretrained language model, and $x'\backslash w$ is x' that w is replaced with a padding token.

Adversarial Domain Adaptation (ADA) ADA is an unsupervised domain adaptation technique (Ganin et al. 2016; Ganin and Lempitsky 2015). In ADA, a model learns to make the features of unlabeled data from a target domain closer to the features of labeled data from a source domain while learning a task using the labeled data. This training is done in an adversarial manner. During training, an additional neural network called discriminator D is trained. The discriminator is given the output of the encoder and predicts 1 if the input is source domain data and 0 otherwise. The encoder learns to fool the discriminator. We use ADA considering the labeled dataset as source domain data and the unlabeled dataset as target domain data. Formally, the objective is written as follows:

$$\begin{aligned}\mathcal{L}_{\text{ADA}} =& \mathbb{E}_{x\sim\mathcal{D}^l}\text{BCE}(0, D(E(x))) \\ &+ \mathbb{E}_{x\sim\mathcal{D}^u}\text{BCE}(1, D(E(x))).\end{aligned} \tag{5}$$

Consistency (CON) CON learns the consistency of volitionality classification and subject animacy classification on the unlabeled datasets. Recall that animacy is a necessary condition for volitionality. Therefore, it is implausible to predict that an event is volitional while predicting that its subject is inanimate. CON learns this relationship by:

$$\mathcal{L}_{\text{CON}} = \mathbb{E}_{x\sim\mathcal{D}^u_{\text{vol}}+\mathcal{D}^u_{\text{ani}}}\max(0, C_{\text{vol}}(E(x)) - C_{\text{ani}}(E(x))). \tag{6}$$

These regularization objectives are combined with the classification objective with a weight. Our training objective is finally written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{WR|SOC|ADA}} + \beta\mathcal{L}_{\text{CON}}, \tag{7}$$

where α and β are weights selected as hyper-parameters and $\mathcal{L}_{\text{WR|SOC|ADA}}$ is either $\mathcal{L}_{\text{WR}}$, $\mathcal{L}_{\text{SOC}}$, or $\mathcal{L}_{\text{ADA}}$.[6]

[6]It is possible to combine WR, SOC, and ADA, in theory. We

Japanese		English	
Volitional	Non-volitional	Volitional	Non-volitional
aete (5,293)	*omowazu* (18,115)	carefully (13,594)	unfortunately (13,070)
isoide (4,187)	*tsui* (15,897)	thoroughly (12,468)	automatically (12,824)
jikkuri (4,017)	*jidoutekini* (14,212)	actively (10,379)	accidentally (5,272)
shinchoni (3,743)	*futo* (12,050)	deliberately (3,366)	unexpectedly (3,106)
wazawaza (3,262)	*tsuitsui* (10,054)	intentionally (2,713)	luckily (1,894)

Table 1: The five most frequent volitionality indicating words in our lexicon. The numbers in parentheses indicate frequency.

	$\mathcal{D}^l_{\text{vol}}$	$\mathcal{D}^u_{\text{vol}}$	$\mathcal{D}^l_{\text{ani}}$	$\mathcal{D}^u_{\text{ani}}$
Japanese	78.2	76.9	77.2	74.3
English	62.4	62.8	66.8	67.1

Table 2: The inter-annotator agreement rate for each dataset, calculated by averaging the ratios of majority answers.

5 Experiments

5.1 Training Dataset

We constructed training datasets following the procedure described in Section 4.1.

Japanese We used 30M documents in CC-100 as a raw corpus (Conneau et al. 2020; Wenzek et al. 2020). Events were parsed and extracted using KNP, a widely used Japanese parser (Kawahara and Kurohashi 2006). For volitionality labeling, we manually constructed a lexicon of 15 volitional and 15 non-volitional adverbs (Table 1). For animacy labeling, we used the dictionary on which KNP builds[7] as ontological knowledge. It contained approximately 30K nouns with animacy labels. We also used the named entity recognizer built into KNP to recognize animacy. We did not delete duplicate events to preserve frequency information.

English We again used 30M documents in CC-100 as a raw corpus. Events were parsed and extracted using spacy[8]. For volitionality labeling, we manually constructed a lexicon of 10 volitional and 10 non-volitional adverbs (Table 1). For animacy labeling, we obtained animate/inanimate nouns from ConceptNet (Speer, Chin, and Havasi 2017). Specifically, we used the hyponyms of "person" and "organization" as animate nouns, and the hyponyms of "object," "item," "thing," "artifact," and "location" as inanimate nouns. As a result, we obtained 2,604 animate nouns and 430 inanimate nouns. Besides, we used the named entity recognizer built into spacy for animacy recognition.

5.2 Evaluation Dataset

We constructed an evaluation dataset for each of $\mathcal{D}^l_{\text{vol}}$, $\mathcal{D}^u_{\text{vol}}$, $\mathcal{D}^l_{\text{ani}}$, and $\mathcal{D}^u_{\text{ani}}$. We first randomly extracted 1,200 unique events from each of the datasets. We then assigned the ground truth to them by crowdsourcing. As for volitionality labeling, crowdworkers were given an event and assigned one of the following labels:

- The subject is volitionally involved in the event.
- The subject is not volitionally involved in the event.
- Unable to say either.
- Unable to understand.

As for animacy labeling, crowdworkers were given an event and assigned one of the following labels:

- The subject is a person(s) or organization(s).
- The subject is neither a person(s) nor organization(s).
- Unable to say either.
- Unable to understand.

Each event was annotated by five crowdworkers. One crowdworker annotated ten events. For Japanese, we used

	Split	Label	Japanese	English
$\mathcal{D}^l_{\text{vol}}$	Train	Volitional	31,812	47,926
		Non-volitional	81,002	40,564
	Dev	Volitional	149	67
		Non-volitional	233	92
	Test	Volitional	149	68
		Non-volitional	233	93
$\mathcal{D}^u_{\text{vol}}$	Train	Unlabeled	112,814+	88,490+
	Dev	Volitional	206	62
		Non-volitional	164	104
	Test	Volitional	206	63
		Non-volitional	164	104
$\mathcal{D}^l_{\text{ani}}$	Train	Animate	29,344+	71,257+
		Inanimate	83,470+	17,233+
	Dev	Animate	175	170
		Inanimate	199	59
	Test	Animate	176	170
		Inanimate	200	60
$\mathcal{D}^u_{\text{ani}}$	Train	Unlabeled	112,814+	88,490+
	Dev	Animate	246	78
		Inanimate	93	152
	Test	Animate	246	78
		Inanimate	93	153

Table 3: Statistics of our dataset. The number with + means that the events were randomly sampled from a larger set according to the size of smallest dataset, $\mathcal{D}^l_{\text{vol}}$.

did not try that due to the computational cost.

[7]https://github.com/ku-nlp/JumanDIC

[8]https://spacy.io

Vol.	Ani.		Japanese				English			
			$\mathcal{D}^l_{\text{vol}}$	$\mathcal{D}^u_{\text{vol}}$	$\mathcal{D}^l_{\text{ani}}$	$\mathcal{D}^u_{\text{ani}}$	$\mathcal{D}^l_{\text{vol}}$	$\mathcal{D}^u_{\text{vol}}$	$\mathcal{D}^l_{\text{ani}}$	$\mathcal{D}^u_{\text{ani}}$
NONE	VAN	+ CON	65.3±2.6	77.3±0.9	92.0±0.7	81.4±0.8	64.0±0.9	69.3±0.7	83.5±1.0	82.4±1.3
	WR	+ CON	73.5±1.4	85.1±1.0	94.3±0.6	86.4±0.2	65.1±0.6	70.7±0.2	84.2±0.4	81.7±0.4
	SOC	+ CON	73.7±2.9	82.3±1.7	93.9±0.2	84.5±1.3	63.5±2.0	70.0±0.6	84.3±1.6	82.7±1.7
	ADA	+ CON	69.7±1.0	81.5±0.7	92.7±0.7	81.9±4.2	62.9±3.6	69.6±1.2	84.6±1.2	83.4±2.0
VAN	NONE	+ CON	91.7±1.0	89.5±1.1	72.6±2.4	70.7±2.7	73.7±1.8	66.2±0.9	67.2±3.6	66.0±1.9
	VAN		91.8±1.3	90.6±0.1	91.3±0.3	81.7±1.7	74.4±0.5	70.7±3.1	82.3±2.1	81.5±1.0
		+ CON	92.1±0.5	89.6±1.9	87.7±3.3	83.0±1.2	74.0±1.6	69.8±3.0	84.3±1.6	81.9±1.3
	WR		93.9±0.6	92.5±1.2	94.0±0.0	86.4±0.4	72.4±4.3	69.7±0.2	83.6±0.7	81.6±0.1
		+ CON	92.1±0.9	92.8±1.0	96.0±0.5	88.5±0.3	71.9±2.6	70.8±0.6	84.1±0.8	81.9±0.6
	SOC		90.7±0.7	94.7±0.7	92.8±1.1	85.4±0.7	72.8±1.9	72.0±0.4	83.5±2.9	78.6±2.9
		+ CON	91.5±1.0	93.5±0.6	89.6±1.5	83.7±0.8	72.8±1.4	69.5±1.1	84.9±0.4	82.3±0.3
	ADA		92.3±0.4	89.9±2.7	87.2±3.3	82.2±2.0	74.4±0.7	72.8±2.2	84.1±1.3	82.8±1.4
		+ CON	92.2±0.4	90.9±3.1	87.7±3.1	82.0±2.1	72.0±1.3	70.9±2.0	82.2±1.7	82.3±1.3
WR	NONE	+ CON	91.5±1.5	91.2±0.8	57.3±10.6	57.6±10.7	69.8±0.8	70.0±0.3	55.5±0.5	67.4±1.1
	VAN		92.4±0.8	91.9±0.1	88.8±7.6	83.9±1.4	73.0±1.0	73.1±2.3	82.6±2.1	84.0±0.5
		+ CON	93.2±0.8	93.2±1.3	84.7±5.8	82.2±1.4	72.3±0.5	72.4±1.2	82.3±0.9	83.7±0.5
	WR		91.8±0.7	93.2±0.9	94.3±1.2	87.1±1.4	72.4±0.8	75.5±1.0	82.5±2.2	83.6±0.2
		+ CON	93.4±1.3	93.0±1.0	93.6±1.5	86.3±1.1	72.6±0.8	71.9±1.4	82.0±0.8	84.5±1.0
	SOC		92.1±0.4	94.9±0.4	88.9±3.8	83.6±2.4	72.9±1.4	75.2±1.4	81.2±2.4	82.8±0.7
		+ CON	93.5±1.3	93.3±1.2	85.0±5.5	82.3±1.4	72.3±1.1	75.5±1.8	80.7±2.2	83.5±1.8
	ADA		92.4±0.5	92.3±1.3	83.6±3.4	81.9±0.3	72.3±0.7	75.6±0.6	82.3±2.1	83.6±0.2
		+ CON	93.9±1.0	93.1±0.8	85.1±5.5	81.7±1.0	72.3±0.8	72.7±0.7	81.4±1.3	83.9±0.4
SOC	NONE	+ CON	94.4±0.6	92.9±0.1	67.3±2.2	67.1±0.7	73.3±0.4	72.2±1.3	66.2±1.8	73.0±1.3
	VAN		94.6±0.4	94.0±0.5	92.3±1.8	86.1±0.8	73.9±0.5	75.4±0.8	83.2±1.9	83.3±0.2
		+ CON	94.6±0.4	94.7±0.5	90.0±1.0	84.5±0.4	73.7±0.3	75.8±0.5	85.4±0.3	**85.1±0.3**
	WR		94.3±0.1	**96.7±0.7**	95.3±1.0	**89.9±0.6**	73.2±0.3	74.0±1.9	84.0±0.1	83.0±0.4
		+ CON	94.5±0.3	**96.7±0.4**	90.1±0.9	84.5±0.6	73.6±0.2	75.7±0.2	84.6±0.9	84.7±0.4
	SOC		94.5±0.3	95.2±0.3	91.3±1.3	86.0±0.8	73.8±0.1	76.7±1.4	**86.2±0.4**	81.4±1.2
		+ CON	94.4±0.4	96.0±0.5	90.0±0.4	84.6±0.8	73.6±0.3	**77.1±1.1**	85.4±0.8	84.4±0.4
	ADA		94.6±0.5	95.9±0.1	92.1±2.0	85.3±0.9	73.1±0.4	74.0±1.8	83.9±0.4	83.1±0.2
		+ CON	95.0±0.8	95.1±1.2	90.4±1.0	84.6±0.8	73.5±0.3	75.4±0.3	84.7±0.8	84.7±0.5
ADA	NONE	+ CON	**96.3±0.6**	93.6±0.9	73.6±2.1	73.4±1.6	**74.9±1.3**	70.0±3.4	62.2±3.2	64.4±1.9
	VAN		90.7±0.4	91.8±0.5	90.8±0.7	82.4±1.7	70.5±2.3	70.5±1.3	84.9±0.6	80.4±1.4
		+ CON	92.1±0.5	89.5±2.4	86.9±4.3	82.8±1.5	72.3±4.4	69.9±1.9	84.9±0.7	81.0±2.6
	WR		92.0±1.3	94.6±0.4	94.9±1.3	86.8±1.1	71.7±0.8	70.6±0.6	83.3±0.7	82.2±1.3
		+ CON	93.1±1.0	94.5±1.0	**95.9±0.3**	87.6±0.8	69.7±2.4	71.5±1.1	85.2±1.0	80.4±2.5
	SOC		91.2±0.6	94.6±0.9	91.2±2.4	84.3±0.5	69.1±3.4	73.2±1.6	84.0±3.0	78.2±1.6
		+ CON	91.6±0.6	93.6±0.2	88.7±1.0	83.3±0.2	69.4±1.7	68.3±0.8	85.1±0.8	82.0±1.7
	ADA		91.9±0.3	90.8±1.1	87.3±2.7	83.1±0.9	71.5±0.2	67.6±2.6	84.0±0.6	77.0±2.3
		+ CON	92.2±0.4	91.3±1.4	87.5±2.9	83.4±0.8	71.5±2.9	71.0±2.8	84.9±0.7	81.0±2.1

Table 4: The result of volitionality classification and subject animacy classification. Reported scores are the average and standard deviation of the AUC of the ROC curve when we trained each model three times with different random seeds. VAN means that classification is learned without regularization (i.e., $\alpha = 0.0$). NONE means that classification is not learned using its labeled data, but learned through optimizing prediction consistency with CON. The bold scores indicate the highest ones over models, and the underlined scores indicate the highest ones over models that did not rely on joint learning.

Yahoo! Crowdsourcing[9] as a crowdsourcing platform. For quality control, we used the function provided in the platform to reject workers who made a mistake on an easy question that we manually prepared in advance. The total cost was 24,000 JPY. For English, we used Amazon Mechanical Turk (MTurk). For quality control, we followed common best practices (Berinsky, Huber, and Lenz 2012); workers had to have over a 95% acceptance rate, live in the US, and have done more than 1,000 tasks. The total cost was 288 USD.

Table 2 shows the inter-annotator agreement rates. Events with an agreement rate of 80% or more were extracted, and half were used for validation and the other half were used for testing. Table 3 summarizes the constructed datasets.

5.3 Implementation Detail

The encoder was pretrained $\text{BERT}_{\text{BASE}}$ (Devlin et al. 2019). We used the output of the classification token ([CLS]) as

[9]https://crowdsourcing.yahoo.co.jp/

event representations. The classifiers were a three-layered fully-connected neural network with the ReLU nonlinearity followed by the sigmoid function. The discriminator used in ADA had the same architecture as the classifiers. SOC was used with a sample size of three. α was selected from $\{0.0, 0.01, 0.1, 1.0\}$ for each of WR, SOC, and ADA. β was selected from $\{0.0, 0.01, 0.1, 1.0\}$. We trained the model for three epochs with a batc/h size of 256. We used the Adam optimizer (Kingma and Ba 2015) with a learning rate of 3e-5, linear warmup of the learning rate over the first 10% steps, and linear decay of the learning rate. We evaluated the performance on the development dataset of $\mathcal{D}^u_{\text{vol}}$, which was our primary concern, at every 100 steps, and adopted the checkpoint that achieved the best performance. The evaluation metric was the AUC of the ROC curve. Models were trained three times with different random seeds. We used Pytorch for implementation.

5.4 Results

Table 4 shows the result. In both Japanese and English, joint learning combined with regularization achieved the best performance on both volitionality and subject animacy classification on the unlabeled datasets and most of the labeled datasets. Specifically, when joint learning was employed, SOC was constantly effective to learn volitionality classification. Without joint learning, the models trained with ADA often performed best.

As for subject animacy classification, the effective method depended on language. This is likely because Japanese is a pro-drop language while English is not. In Japanese, the most effective method was WR. Learning subject animacy of events produced by WR can be interpreted as learning animacy of omitted subjects. Events with omitted subjects were not given a subject animacy label by our labeling method and thus were in $\mathcal{D}^u_{\text{ani}}$. The models trained with WR successfully generalized to such events. In English, on the other hand, because subjects are not omitted, WR was not as effective as in Japanese.

We observed that the overall scores on the English datasets were lower than the Japanese ones. The reason was the quality of the evaluation datasets. As Table 2 suggests, the English evaluation datasets were constructed by crowdworkers with a lower agreement rate. Investigating the output manually, we found that the performance was underestimated due to labeling mistakes in the gold data.

6 Analysis

6.1 Qualitative Analysis

We investigated what is learned by our method, using the model that best performed on $\mathcal{D}^u_{\text{vol}}$. While we had three models trained with the same setting with different random seeds, we used one that achieved the second-best validation performance for analysis.

Japanese The best performing model learned volitionality with SOC, subject animacy with WR, and prediction consistency with CON. We found that the model was aware of context. Example (1-a) and (1-b) were successfully classified as volitional and non-volitional, respectively, though these events had the same predicate. Example (2-a) and (2-b) were again correctly classified as non-volitional and volitional, respectively, considering the meaning of the adverb. We observed that subject animacy was also recognized considering context; the subjects of example (3-a) and (3-b) were successfully classified as inanimate and animate, respectively. It would be interesting to quantitatively evaluate such context-awareness by constructing a dataset like Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2012).

However, we found that there still existed verbs that our model struggled with recognizing the volitionality. One notable verb was "*iru* (exist/stay)." While the verb "*iru* (exist/stay)" basically represents a state, it can represent a volitional action when the subject is animate. We speculate that the difficulty of recognizing the animacy of omitted subjects also contributed to this problem. A plausible solution is to consider the preceding and following events during training. If the meaning of an event is different, the distribution of its surrounding events should be too. Learning such contextual differences could lead to better performance.

English The best performing model learned volitionality with SOC, subject animacy with SOC, and prediction consistency with CON. We again found that the model successfully performed classification considering context. For example, the following examples with the same predicate "made" were correctly classified.

(13) a. I made pancakes. (V)
b. I made a mistake. (NV)

The following examples were also successfully classified, capturing the meaning of the adverbial phrase "for him."

(14) a. I tumbled. (NV)
b. I tumbled for him. (V)

	Label	Japanese	English
$\mathcal{D}^l_{\text{vol}}$	Volitional	88%	94%
	Non-volitional	92%	80%
$\mathcal{D}^l_{\text{ani}}$	Animate	81%	96%
	Inanimate	72%	76%

Table 5: The ratio of events being given a correct label.

6.2 Quality of Labeled Data

Because we had heuristically and automatically assigned labels to events, our labeled datasets should contain wrongly labeled events. However, if the datasets were full of errors, it is likely to fail to learn classification.

Given the fact that we could learn a classifier with fairly good performance, we report the quality of our labeled data as a reference for applying our method to other languages. We randomly extracted 100 unique positive and negative events from each of $\mathcal{D}^l_{\text{vol}}$ and $\mathcal{D}^l_{\text{ani}}$, and manually examined whether they were given a correct label or not. We considered that events incomprehensible for some reason (e.g.,

parsing error) were not given a correct label.

Table 5 shows the result. We found that most events were labeled correctly. Japanese negatively-labeled events in $\mathcal{D}^l_{\text{ani}}$ had relatively low accuracy. This was primarily because of the failure in subject recognition. In English, negatively-labeled events in $\mathcal{D}^l_{\text{ani}}$ had relatively low accuracy. While there were several reasons, one of them was that, although we regarded nouns representing a location as inanimate, they sometimes represented an organization (e.g., country name).

7 Conclusion

This paper focused on the close relationship between volitionality and animacy and proposed a method to jointly learn them with regularization in a minimally-supervised manner. Experiments in Japanese and English showed the effectiveness of the proposed method to learn volitionality and subject animacy without manually labeled data.

Acknowledgements

This work was partly supported by Yahoo Japan Corporation.

References

Abe, S.; Inui, K.; and Matsumoto, Y. 2008a. Acquiring Event Relation Knowledge by Learning Cooccurrence Patterns and Fertilizing Cooccurrence Samples with Verbal Nouns. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Abe, S.; Inui, K.; and Matsumoto, Y. 2008b. Two-Phased Event Relation Acquisition: Coupling the Relation-Oriented and Argument-Oriented Approaches. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 1–8. Manchester, UK: Coling 2008 Organizing Committee.

Berinsky, A.; Huber, G.; and Lenz, G. S. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20: 351 – 368.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bowman, S. R.; and Chopra, H. 2012. Automatic Animacy Classification. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, 7–10. Montréal, Canada: Association for Computational Linguistics.

Chambers, N.; and Jurafsky, D. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, 789–797. Columbus, Ohio: Association for Computational Linguistics.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, L.; Ding, X.; Liu, T.; and Li, Z. 2019. Modeling Event Background for If-Then Commonsense Reasoning Using Context-aware Variational Autoencoder. 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)*, 2682–2691. Hong Kong, China: Association for Computational Linguistics.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised Domain Adaptation by Backpropagation. In Bach, F.; and Blei, D., eds., *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 1180–1189. Lille, France: PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59): 1–35.

Gildea, D.; and Jurafsky, D. 2000. Automatic Labeling of Semantic Roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 512–520. Hong Kong: Association for Computational Linguistics.

Inui, T.; Inui, K.; and Matsumoto, Y. 2003. What kinds and amounts of causal knowledge can be acquired from text by using connective markers as clues? In *International Conference on Discovery Science*, 180–193. Springer.

Jin, X.; Wei, Z.; Du, J.; Xue, X.; and Ren, X. 2020. Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models. In *International Conference on Learning Representations*.

Kawahara, D.; and Kurohashi, S. 2006. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 176–183. New York City, USA: Association for Computational Linguistics.

Kennedy, B.; Jin, X.; Mostafazadeh Davani, A.; Dehghani, M.; and Ren, X. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5435–5442. Online: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*.

Lee, H.-Y.; and Jun, S. 2008. Constructing an Ontology of Coherence Relations: An example of 'causal relation'. In

Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, 245–252. The University of the Philippines Visayas Cebu College, Cebu City, Philippines: De La Salle University, Manila, Philippines.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, 552–561. AAAI Press. ISBN 9781577355601.

Liu, C.-H.; Moriya, Y.; Poncelas, A.; and Groves, D. 2017. IJCNLP-2017 Task 4: Customer Feedback Analysis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, 26–33. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Ramponi, A.; and Plank, B. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6838–6855. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Reisinger, D.; Rudinger, R.; Ferraro, F.; Harman, C.; Rawlins, K.; and Van Durme, B. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3: 475–488.

Shah, D.; Lei, T.; Moschitti, A.; Romeo, S.; and Nakov, P. 2018. Adversarial Domain Adaptation for Duplicate Question Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1056–1063. Brussels, Belgium: Association for Computational Linguistics.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein Distance Guided Representation Learning for Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451. AAAI Press.

Teichert, A.; Poliak, A.; Van Durme, B.; and Gormley, M. 2017. Semantic Proto-Role Labeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; and Grave, E. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4003–4012. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

White, A. S.; Reisinger, D.; Sakaguchi, K.; Vieira, T.; Zhang, S.; Rudinger, R.; Rawlins, K.; and Van Durme, B. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723. Austin, Texas: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. Minneapolis, Minnesota: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics.