

Extracting Possessions from Social Media: Images Complement Language

Dhivya Chinnappa, Srikala Murugan and Eduardo Blanco

Human Intelligence and Language Technologies Lab

University of North Texas

Denton, Texas 76203

{dhivyainfantchinnappa, srikalamurugan}@my.unt.edu, eduardo.blanco@unt.edu

Abstract

This paper describes a new dataset and experiments to determine whether authors of tweets possess the objects they tweet about. We work with 5,000 tweets and show that both humans and neural networks benefit from images in addition to text. We also introduce a simple yet effective strategy to incorporate visual information into any neural network beyond weights from pretrained networks. Specifically, we consider the tags identified in an image as an additional textual input, and leverage pretrained word embeddings as usually done with regular text. Experimental results show this novel strategy is beneficial.

1 Introduction

Social media are platforms for sharing information online. Social media posts and online behavior in general (e.g., Facebook likes, following other users in Twitter) have been shown to predict human traits (Burke et al., 2010; Schwartz et al., 2013). Many social media posts include an image alongside text, and the percentage keeps growing as doing so boosts user engagement (Patel, 2016).

Pictures and text in social media usually complement each other. Thus, even if the information of interest can be understood from either communication modality, considering both is beneficial. Consider the tweet in Figure 1. The text indicates that *Arnold* (the author of the tweet) goes on *bike* rides when he travels. The image shows him riding a bike, indicating that he was riding a *bike* when he tweeted thus he was in possession of a *bike*. On the other hand, if the picture were a screenshot of his Twitter posting statistics, *Arnold* most likely would not be in possession of a *bike* when tweeting, but rather sharing a log of his previous trips with his followers. In this paper, we extract possession relations from social media



Figure 1: Sample tweet with text and an image. We can infer that *Arnold* possessed a *bike* when he tweeted.

posts containing both text and images. Possession is an asymmetric semantic relation between two entities, where one entity (the possessee) *belongs* to the other entity (the possessor) (Stassen, 2009). Following the literature, we consider not only ownership, but also control possessions. In control possessions, the possessor has temporary control of the possessee but not necessarily ownership (Tham, 2004), e.g., *Bill borrowed the ozone generator from John*.

While we do not explore any, extracting possessions has many potential applications. For example, possessions could help to reveal hobbies and to find people with similar interests. Possessions could also improve recommender systems. For example, people without cars are unlikely to be interested in oil changes and auto mechanics. Similarly, people who recently purchased a home may be interested in moving and remodeling services. Extracting possessions could also be useful to identify skills. For example, peo-

ple who possess a bike are likely to be able to ride bikes, and those who have control possession of an 18-wheeler are typically able to drive large trucks. The main contributions of this paper are: (a) a corpus of 5,000 tweets (text and images) annotated with possession relations including type (alienable or control), temporal anchors with respect to the tweet timestamp, and interest in something,¹ (b) detailed corpus analysis showing, among others, that humans understand more possession relations when they have access to both the text and images, (c) experimental results showing that the task can be automated and features extracted from the images improve results. Regarding visual features, we show that incorporating weights from pretrained networks—a common practice in previous work—is beneficial, but we obtain more substantial improvements incorporating the objects and events identified in an image as an additional textual input and leveraging word embeddings.

2 Previous Work

Possession relations have primarily been studied in efforts targeting large relation repositories between arguments connected with some lexico-syntactic pattern. [Tratz and Hovy \(2013\)](#) work with 17 semantic relations realized by possessive constructions, [Badulescu and Moldovan \(2009\)](#) with 36 relations realized by genitives, and [Nakov and Hearst \(2013\)](#) and [Tratz and Hovy \(2010\)](#) target noun compounds. [Blodgett and Schneider \(2018\)](#) annotate 50 supersenses (including roles and relations between entities) for possessives. These efforts extract possessions from text, and target possessors and possessees connected by specific patterns. Unlike them, we extract possessions using both text and images. In addition to possession existence, we also extract types, temporal anchors, and interest in the possessee.

Two recent efforts target possession relation extraction from text without strict syntactic constraints. In our previous work, we extract intra-sentential possessions from OntoNotes ([Chinnappa and Blanco, 2018](#)). In the work described here, we use the list of synsets from our previous work to select possessees (Section 3). [Banea and Mihalcea \(2018\)](#) work with blogs and annotate possession existence at the time of utterance. Unlike these previous works, we (a) leverage both

text and images, (b) work with informal tweets (instead of standard English), (c) temporally anchor possessions before, during and after the tweet timestamp, and (d) also extract whether somebody has an interest in a concrete object regardless of possession existence.

Using multiple modalities (e.g., text and images) to better solve some task is not new. Among many others, [Specia et al. \(2016\)](#) propose multimodal machine translation and [Moon et al. \(2018\)](#) show that named entity recognition benefits from taking into account both text and images. Our innovation is twofold. First, we show that humans understand more possession information when they have access to the image accompanying a text, as opposed to only reporting improvements on (automatically) solving some task. Second, our neural image component includes two subcomponents. The first one—weights from InceptionNet—is common in previous work, but the second one is novel. Specifically, the second component considers the objects and events identified in an image as an additional textual input. This allows us to leverage pretrained word embeddings and recurrent neural networks, a strategy that we prove beneficial.

3 A Corpus of Possession Relations

We start with a collection of English tweets consisting of text and images ([Hu et al., 2018](#)). First, we discard tweets that do not contain *I*, *me*, *my*, or *mine* in order to maximize the amount of tweets published by individuals and avoid tweets by organizations as well as advertisements. Second, we select as potential possessors the authors of tweets, and as potential possessees the nouns subsumed by the WordNet synsets ([Miller, 1995](#)) proposed in previous work (Section 2) except the following nouns: *fan*, *filter*, *launch*, *mini*, *release* and *safe*. We eliminate them because they almost never yield possession relations in social media. For example, *fan* almost always refers to a person (e.g., *This Bucks fan put on a show*) instead of to an “apparatus with rotating blades,” and *filter* almost always refers to a photo effect (e.g., *Bare face plus a snap filter*) instead of to a “porous device for removing impurities.” Finally, we randomly select 5,000 possessor-possessee pairs.

3.1 Annotation Tasks And Guidelines

In addition to possession existence (i.e., whether the potential possessor possesses the potential pos-

¹Available at dhivyachinnappa.com

	Only Text		Text and Image	
	%	κ	%	κ
alienable, control, never, unk	87.1	0.83	86.5	0.82
before yes, before no	90.0	0.80	91.6	0.81
during yes, during no	89.5	0.78	98.1	0.78
after yes, after no	90.5	0.80	97.2	0.79
interest_yes, interest_no	82.5	0.76	90.3	0.78

Table 1: Inter-annotator agreements (observed and Cohen’s κ) having access to (a) the text and (b) the text and image. κ values between 0.60 and 0.80 are considered *substantial*, and above 0.80 *nearly perfect*.

sessee), we also annotate possession type (alienable or control), temporal anchors with respect to the tweet timestamp (before, during and after), and whether the potential possessor has an interest in the potential possessee regardless of possession existence and possession type.

Possession Existence. The first annotation task is to determine whether the potential possessor (x) possesses the potential possessee (y). Annotators choose between the following labels:

- *yes* if a possession exists (i.e., x possesses y) at some point of time;
- *never* if a possession does not exist (i.e., x does not possess y) at any point of time; or
- *unk* (unknown) if it is sound to ask whether x possesses y , but there is not enough information to choose *yes* or *never*.

Possession Type. If a possession relation exists (existence: *yes*), annotators also indicate the type:

- *alienable*: if x can be separated from y and x is the owner of y , regardless of spatial proximity or other variables; or
- *control*: if x can be separated from y and x has control over y , regardless of ownership, spatial proximity or other variables.

Note that according to these definitions, control possessions, unlike alienable possessions, do not require ownership. For example, people driving a rental car have control possession of the car but not alienable possession. Control possession and alienable possession are mutually exclusive labels. We study possession types (*alienable* and *control*) to understand the strength of the possession relation between the possessor and the possessee. We do not consider inalienable possessions because they are uncommon in social media.

Temporal Anchors. If a possession relation exists (existence: *yes*), annotators also indicate when it is true with respect to the tweet timestamp:

- *before yes* or *no*: whether x possesses y the

day before tweeting or earlier;

- *during yes* or *no*: whether x possesses y the day he tweeted; and
- *after yes* or *no*: whether x possesses y the day after tweeting or later.

Interest in the Possessee. Finally, annotators also indicate whether x has an interest in y (*interest_yes* or *interest_no*), regardless of the labels for possession existence and type. Interest does not entail past, current or future possession existence. It indicates that x shows curiosity or excitement about y . Let us consider John Doe and a tweet about eating more vegetables (with a fork) because of a doctor’s recommendation. In this context, John would have possession of the *fork* but no interest in it.

3.2 Annotation Set-Up and Quality

Annotations were done in-house by two graduate students. We developed a simple interface that showed one tweet at a time, and annotators were instructed to use world knowledge and common sense. In order to minimize biases, the interface did not show the twitter handle, profile picture or any other user information. In addition, we sampled 200 tweets and only in 7% of them the image possibly provided information about the tweet author (e.g., gender, age, race, ethnicity).

In a first round of annotations, annotators had access only to the text in the tweet. In a second round, they had access to both the text and the image. Our rationale is that we are interested in comparing human judgments depending on whether the image is available or not (Section 4.1).

Inter-Annotator Agreement. Table 1 shows inter-annotator agreements (observed and Cohen’s κ) when annotators have access to (a) only the text and (b) the text and image. Agreements are very similar regardless of whether the image is available, and as we shall see (Section 4.1), doing so results in more possession information. Cohen’s

Possession existence and type	yes, alienable	38.4%
	yes, control	7.5%
	never	38.6%
	unk	15.5%
Interest	interest_yes	40.2%
	interest_no	59.8%

Table 2: Label distributions. We divide the percentage of *yes* into *alienable* and *control*.

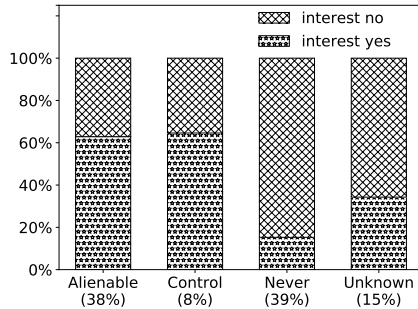


Figure 2: Distribution of *interest in the possessee* labels depending on possession existence and type. People often have an interest in objects they do not possess and when possessions cannot be determined (*never*, *unk*).

κ for possession existence and type (first row) is 0.82 with text and images, it ranges from 0.78 to 0.81 for temporal anchors (rows 2–4), and it is 0.78 for interest in the possessee (row 5). κ values between 0.60 and 0.80 are *substantial*, and above 0.80 *nearly perfect* (Artstein and Poesio, 2008).

4 Corpus Analysis

We start describing the final corpus, which was annotated with both text and images. Then, we compare the corpus obtained when annotators have access to (a) only the text and (b) the text and image. **Label Distributions.** Table 2 shows the label distributions for the three annotation tasks: possession existence, possession type (*yes*: *alienable* or *control*) and interest in the possessee.

Overall, the percentage of unknown label (*unk*) is low (15.5%), indicating that possession existence can almost always be determined. More importantly, the procedure described in Section 3 allows us to reveal useful knowledge in 84.5% of all generated possessor-possessee pairs (*alienable*, *control* and *never*). Most possessions are alienable (83.6%, 38.4% of all possessor-possessee pairs) and the percentage of control possession is low. The *never* label is somewhat common (38.6), indicating that people often tweet about objects that they have never possessed.

		alienable	control
before	yes	79.5%	45.8%
	never	20.5%	54.2%
during	yes	95.9%	76.3%
	no	4.1%	23.7%
after	yes	90.5%	29.5%
	no	9.5%	70.5%

Table 3: Distribution of temporal anchor labels depending on the possession type (*alienable* or *control*).

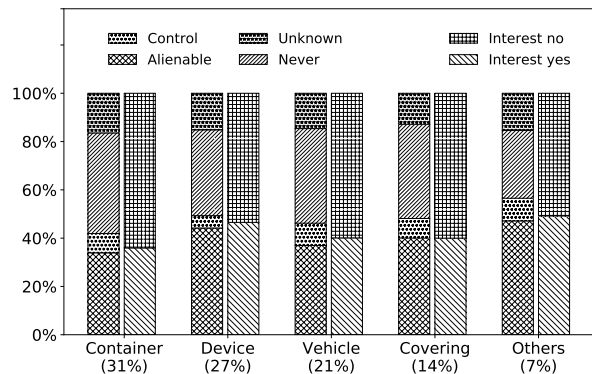


Figure 3: Label distribution (left column: possession existence and type, right column: interest) depending on the WordNet synset of the possessee.

Regarding interest, the possessor has an interest in the possessee in 40.2% of the 5,000 generated pairs. Figure 2 provides insights regarding possession existence and interest for the generated pairs. First, the possessor has an interest in the possessee in (a) 35% of pairs for which possession could not be determined (*unk*) and (b) 15% of pairs for which no possession exists (*never*). Second, regardless of possession type, the percentage of *interest_yes* remains at approximately 60%.

The distributions of temporal anchor labels (*before*, *during* and *after*) per possession type (Table 3) show that possession type substantially influences *when* the possession is true with respect to the tweet timestamp. Regarding alienable possessions, people tweet more about what they own or will own in the future than what they owned in the past (95.9% and 90.5% vs. 79.5%). Control possessions show mostly uniform distribution with anchor *before*, and are unlikely to be true the day after tweeting (29.5%).

Finally, Figure 3 presents the distribution of possession existence, possession type and interest in the possessee depending on the WordNet synset of the possessee. Regarding possession existence and possession type labels (left columns), we observe similar distributions across synsets, although




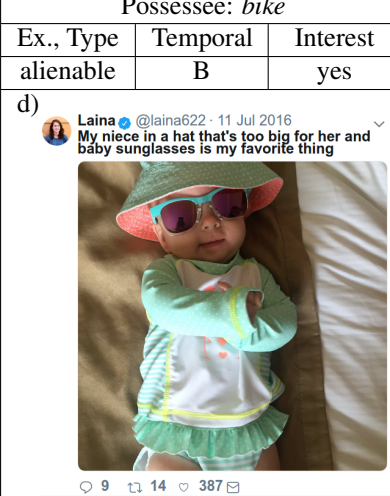
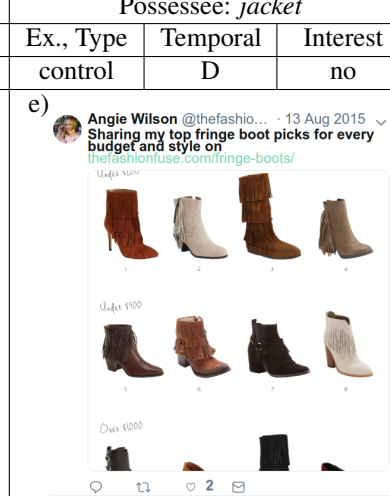
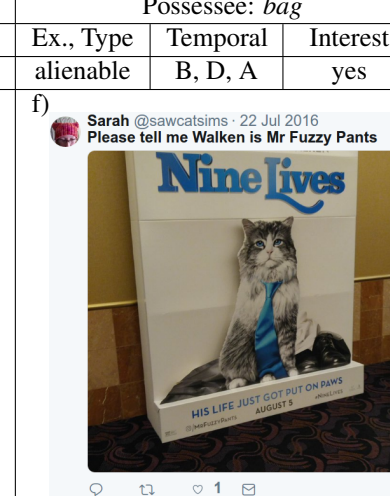
<p>a) </p>			<p>b) </p>			<p>c) </p>		
Possessee: <i>bike</i>			Possessee: <i>jacket</i>			Possessee: <i>bag</i>		
Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest
alienable	B	yes	control	D	no	alienable	B, D, A	yes
<p>d) </p>			<p>e) </p>			<p>f) </p>		
Possessee: <i>sunglasses</i>			Possessee: <i>boot</i>			Possessee: <i>pants</i>		
Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest
never	n/a	no	unk	n/a	yes	never	n/a	no

Table 4: Annotation examples when annotators have access to the text and image. We indicate possession existence and type (alienable, control, never, unk), temporal anchors (B: before, D: during, A: after) and interest.

devices (e.g., watch, guitar, cell phone) yield more possessions (alienable and control labels) and most of the possessions are alienable. In other words, people tend to tweet about *devices* they own, and rarely about *devices* they only have control over. Regarding interest in the possessee (right columns), people are slightly more likely to have an interest if the possessee is a *device*.

Annotation Examples. We present annotation examples in Table 4. Note that unlike in these examples, the annotation interface did not show the Twitter handle and profile picture in an effort to minimize potential biases (Section 3.2).

In Example (a), annotators understood that the author of the tweet was a competitive biker (bike stunt, racing flag), and world knowledge tells us that competitive bikers own their bikes. Thus, the author had an alienable possession with the bike. The text clearly indicates that the possession was

true in the past (*2 years ago*), and hints that the author has an interest in bikes (*miss being on X*).

Example (b) is a straightforward example of control possession: the author does not own the *jacket*. While weekends last for two days, it is unknown when the author tweeted, so annotators chose only *during* temporal anchor. Additionally, neither the text or image indicate that the author has any interest in the *jacket*.

Example (c) illustrates an alienable possession in which the author possesses the possessee (i.e., the *bag*) before, during and after tweeting. While there is no specific cue indicating that the author will own the *bag* for an extended period of time, common sense indicates so. Additionally, the text (*my cutest bag ever*) indicates that the author is excited and has an interest in the *bag*.

Example (d) illustrates *never* label. In this case, the author is talking about the baby's *sun-*

		Only Text			
		alienable	control	never	unk
Text and Image	alienable	80.5%	7.8%	10.0%	38.7%
	control	9.4%	82.8%	3.0%	12.2%
	never	8.3%	6.3%	83.3%	23.4%
	unk	1.8%	3.1%	4.7%	25.7%
Total		100.0%	100.0%	100.0%	100.0%

Table 5: Changes in labels depending on whether annotators have access to the image. Note that 74.3% of instances annotated unk with only text become *alienable* (38.7%), *control* (12.2%) or *never* (23.4%).




	(a)			(b)			(c)		
									
	Possessee: <i>denim</i>			Possessee: <i>candle</i>			Possessee: <i>hat</i>		
	Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest	Ex., Type	Temporal	Interest
T	unk	n/a	no	unk	n/a	no	alienable	B, D, A	no
T+I	alienable	B, D, A	yes	alienable	B, D, A	no	never	n/a	no

Table 6: Examples of tweets which are annotated different depending on whether annotators have access to only the text (first row of labels, T) or the text and image (second row of labels, T+I). Note that the image allows to annotate more possessions (Examples (a) and (b)) as well as fix mistakes (Example (c)).

glasses. Additionally, there is no indication of the author having an interest about the *sunglasses*.

Examples (e, f) illustrate *unk* and *never* labels. The author of tweet (d) is sharing his favorite *boots*, and there is not enough information to determine whether she owns any. The author is, however, interested in *boots*, as she went through the task of choosing her favorite *boot* picks. Finally, in Example (f), *pants* cannot be a possessee as it is part of the name of a movie character.

4.1 Text vs. Text and Image

Annotators chose different labels depending on whether they had access to (a) only the text or (b) the text and image. Table 5 summarizes the changes in annotations. Most labels remain the same (*alienable*: 80.5%, *control*: 82.8%, *never*: 83.3%), however, most instances labeled *unk* when annotators have access only to text (74.3%) become *alienable* (38.7%), *control* (12.2%), or *never* (23.4%) when they also have access to the image (last column).

Table 6 shows examples of changes in annotation. Given only the text in Tweet (a), it appears that the author dislikes *denim*. Looking at the image, however, it becomes clear that the author is being sarcastic and not only owns *denim* clothes but actually has a strong interest in *denim*. The text in Tweet (b) is basically two quotes, and looking only at the text one cannot determine whether the author owns any candles. Taking into account the picture, however, one can conclude that the author does own a *candle* (the picture illustrates the advice from the quote) although she does not have an interest in it. The last example, Tweet (c), illustrates how images also help discarding possessions that appear obvious from the text. The *hat* is not an actual object (it is a drawing on top of the picture) thus no *alienable* possession exists.

5 Experiments and Results

We experiment primarily with neural networks. Regarding libraries, we use Keras (Chollet et al., 2015) with TensorFlow as a backend (Abadi et al.,

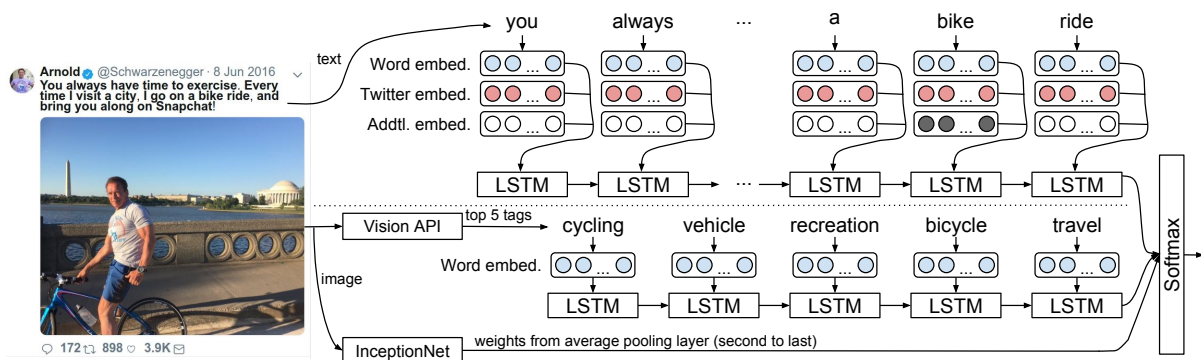


Figure 4: Neural network architecture to predict possession existence, type, temporal anchors and interest. We include a text component (above dotted line) and two image components (below dotted line). Note that the top 5 tags from the Vision API become a textual input, and we use pretrained word embeddings and an LSTM for them.

2015). Each possessor-possessee pair (and corresponding tweet) becomes an instance, and we create stratified training (80%) and test (20%) sets. We train the neural network for up to 200 epochs using the Adam optimizer (Kingma and Ba, 2014), categorical cross entropy, and batch size 32. We stop the training process before 200 epochs if no improvement occurs in the validation set (15% of the training set) for 10 epochs. More specifically, we train six classifiers. The first classifier predicts possession existence (*yes*, *never* or *unk*). The second classifier predicts possession types, i.e., classifies pairs between which a possession exists (*yes*) into *alienable* or *control*. The third, fourth and fifth classifiers predict temporal anchors, i.e., classify pairs between which a possession holds—either *alienable* or *control*—into *before yes* or *before no*, *during yes* or *during no*, and *after yes* or *after no*. Finally, the sixth classifier predicts interest in the possessee (*interest_yes* or *interest_no*).²

5.1 Neural Network Architecture

Figure 4 shows the neural network architecture, which includes components for the text and image (above and below dotted line respectively).

Text Component. The text component is an LSTM (Hochreiter and Schmidhuber, 1997). Each token is represented with the concatenation of three embeddings. The first two are GloVe word embeddings pretrained with Common Crawl and Twitter (Pennington et al., 2014). The third embedding only takes two possible values (dark grey: possessee, white: non-possessee) and it is used to indicate the potential possessee. Only the addi-

tional embeddings are tuned along with other network parameters. Intuitively, the additional embedding allows the LSTM to focus on the context surrounding the potential possessee.

Image Component. The image component leverages two pretrained neural networks: InceptionNet (Szegedy et al., 2015) and Cloud Vision API.³ Generally speaking, InceptionNet is pretrained to identify objects, and the Vision API outputs tags describing images including not only objects but also events (e.g., *cycling*, *recreation*, *travel* from the image in the tweet in Figure 4). Regarding InceptionNet, we follow previous work (Section 2) and include the weights of the average pooling layer (second to last layer). This incorporates features (real numbers) capturing characteristics of the image to the output Softmax layer, where the features are useful for object prediction.

More interestingly, we also incorporate the top 5 tags identified in the image by the Cloud Vision API. The main novelty of our architecture is the strategy to incorporate them. Rather than using one-hot encodings or training special-purpose embeddings, we consider the top 5 tags as an additional textual input and leverage pretrained GloVe word embeddings and an LSTM. Word embeddings allow us to bring meaning to the tags. Intuitively, this is more beneficial than incorporating weights from InceptionNet because the embeddings are a distributed representation of word meaning, and they are useful to, among others, determining word similarity and solving analogies (Pennington et al., 2014). The LSTM is useful because tags may be more than one token (e.g., *whipped cream*, *electronic device*) thus the top 5

²Code available at dhivyachinnappa.com

³<https://cloud.google.com/vision/>

	Maj. baseline			NN only text			NN text + IN			NN text + Itags			NN text + img		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
yes	.46	1.0	.63	.70	.69	.69	.70	.72	.71	.70	.78	.74	.73	.78	.76
never	.00	.00	.00	.57	.55	.56	.57	.57	.57	.62	.56	.59	.64	.63	.63
unk	.00	.00	.00	.57	.55	.56	.51	.47	.49	.64	.54	.59	.64	.53	.58
Macro Avg.	.15	.33	.21	.61	.60	.60	.59	.59	.59	.65	.63	.64	.67	.65	.66
alienable	.84	1.0	.91	.84	.94	.89	.84	.95	.89	.84	.96	.90	.83	.92	.88
control	.00	.00	.00	.19	.07	.10	.21	.07	.10	.29	.09	.14	.75	.82	.77
Macro Avg.	.52	.50	.46	.52	.51	.50	.53	.51	.50	.57	.53	.52	.79	.87	.83
interest.yes	.59	1.0	.74	.54	.39	.45	.52	.48	.50	.52	.43	.47	.52	.43	.47
interest.no	.00	.00	.00	.64	.77	.70	.65	.68	.67	.64	.72	.68	.64	.72	.68
Macro Avg.	.30	.50	.37	.61	.58	.58	.59	.58	.59	.58	.58	.58	.58	.58	.58

Table 7: Results for predicting possession existence, possession type and interest in possessee. We report results with the full network (last column, NN text + img) as well as with selected components: only text, text + weights from InceptionNet (IN), and text + LSTM encoding of top-5 tags detected in the image (text + Itags).

		Before			During			After		
		P	R	F1	P	R	F1	P	R	F1
NN, only text	yes	0.74	0.96	0.83	0.92	0.98	0.95	0.82	0.88	0.85
	no	0.35	0.07	0.11	0.00	0.00	0.00	0.29	0.21	0.24
	Macro Avg.	0.55	0.52	0.47	0.46	0.49	0.48	0.56	0.55	0.55
NN, text + img	yes	0.70	0.78	0.74	0.88	0.97	0.92	0.84	0.89	0.87
	no	0.48	0.38	0.43	0.25	0.08	0.12	0.53	0.41	0.46
	Macro Avg.	0.59	0.58	0.59	0.57	0.53	0.52	0.69	0.65	0.67

Table 8: Results for predicting temporal anchors with the neural network (only text, and text and image).

tags are a sequence of at least five tokens—not necessarily five tokens. To the best of our knowledge, this strategy to incorporate information extracted from an image is novel. While simple, we show that it is effective at determining possession existence despite our dataset is relatively small.

5.2 Experimental Results

Tables 7 and 8 present the experimental results (Precision, Recall and F1-score). We present results per label and the macro average.

Baselines. We use the majority baseline and logistic regression using bag-of-words features (not shown, F1-scores are 0.48 (existence), 0.56 (types) and 0.58 (interest)). The full LSTM (NN text + img) outperforms the baselines predicting possession existence and types (existence F1: 0.66 vs. 0.21–0.64; types F1: 0.83 vs. 0.46–0.52), but all models except the majority baseline perform roughly the same predicting interest in the possessee (F1: 0.58–0.59).

5.2.1 Neural Network

Table 7 presents the results obtained with four versions of the neural network: using (a) only the text component, (b) the text component and the weights from InceptionNet (text + IN), (c) the text

component and the tags from the Vision API as an additional textual input (text + Itags), and (d) the full network (text + img). We also obtained results with only the image components, but do not report the results because they were much worse.

Possession Existence. All variations of the neural network outperform the baselines (logistic regression obtains 0.48 F1, not shown). Weights from InceptionNet do not bring any improvement by themselves, but the tags from the Vision API used as an additional textual input do (F1: 0.64 vs. 0.60). More importantly, combining both of them yields 10% improvement (F1: 0.66 vs. 0.60). Further examination revealed that this is due to leveraging pretrained word embeddings with the tags from the Vision API—using one-hot encodings does not bring improvements (not shown).

Possession Type and Interest in the Possessee. Regarding possession type, we observe a similar trend than with possession existence. This time, however, the differences in results are larger (F1: 0.50 vs. 0.83) and the network with both image components (NN text + img) is the only model predicting `control` reliably (0.77 vs. 0.10–0.14).

Regarding interest in the possessee, all models but the majority baseline (including logistic re-

gression) obtain similar F1s (0.58–0.59). While there is certainly room for improvement, the current results lead to the conclusion that a few keywords are sufficient to obtain 0.58 F1: neither images nor word embeddings bring improvements.

Temporal Anchors. Table 8 presents results obtained with the neural network when predicting temporal anchors. The image components are beneficial with all anchors, especially *before* (F1: 0.47 vs. 0.59, +25%) and *after* (0.55 vs. 0.67, +22%), and to a lesser degree *during* (0.48 vs. 0.52; 8%). F1 scores are higher for *yes* label than *no* label across all temporal anchors.

6 Conclusions

We have presented a corpus of 5,000 tweets and experimental results to extract possession relations. Specifically, we work with text and images in order to reveal the possessors of the author of a tweet. Beyond possession existence, we also consider possession type, temporal anchors with respect to the tweet timestamp, and whether the author has an interest in the potential possessor regardless of possession existence.

The corpus analysis shows that humans understand more possessions when they have access to both the text and images. Authors of tweets often have an interest in potential possessors when there is no possession relation or there is not enough information to determine whether a possession exists (*never* and *unk* labels). Finally, experimental results show that incorporating pretrained networks for object identification and image understanding complement neural components that consider text. Crucially, we show that considering the top 5 tags identified in images (objects and events) as an additional textual input and leveraging word embeddings and recurrent neural networks yields better results than incorporating only weights from intermediate layers, as previous work does.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasude-

van, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Adriana Badulescu and Dan Moldovan. 2009. A semantic scattering model for the automatic interpretation of english genitives. *NLE*.

Carmen Banea and Rada Mihalcea. 2018. *Possession identification in text*. *Natural Language Engineering*, 24(4):589610.

Austin Blodgett and Nathan Schneider. 2018. Semantic Supersenses for English Possessives. In *LREC*.

Moira Burke, Cameron Marlow, and Thomas Lento. 2010. *Social network activity and social well-being*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 1909–1912, New York, NY, USA. ACM.

Dhivya Chinnappa and Eduardo Blanco. 2018. Mining possessions: Existence, type and temporal anchors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 496–505, New Orleans, Louisiana, USA. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. <https://github.com/fchollet/keras>.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. *emoji2vec: Learning emoji representations from their description*. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2018. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Trans. Multimedia*, 20(4):927–938.

Diederik P. Kingma and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*. *CoRR*, abs/1412.6980.

George A. Miller. 1995. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. *Multimodal named entity recognition for short social media posts*. In *Proceedings of the*

- 2018 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860. Association for Computational Linguistics.
- Preslav I. Nakov and Marti A. Hearst. 2013. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Trans. Speech Lang. Process.*, 10(3):13:1–13:51.
- Niel Patel. 2016. 3 Simple Tools To Double Your Twitter Engagement This Week. <https://www.forbes.com/sites/neilpatel/2016/09/26/3-simple-tools-to-double-your-twitter-engagement-this-week/>. [Online; accessed Dec 8th, 2018].
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. *Glove: Global vectors for word representation*. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. *Personality, gender, and age in the language of social media: The open-vocabulary approach*. *PLOS ONE*, 8(9):1–16.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. *A shared task on multimodal machine translation and crosslingual image description*. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553. Association for Computational Linguistics.
- L. Stassen. 2009. *Predicative Possession*. Oxford Studies in Typology and Linguistic Theory. OUP Oxford.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Shiao Wei Tham. 2004. *Representing Possessive Predication: Semantic Dimensions and Pragmatic Bases*. Ph.D. thesis, Stanford University.
- Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 678–687, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Stephen Tratz and Eduard H. Hovy. 2013. Automatic interpretation of the english possessive. In *ACL (1)*, pages 372–381. The Association for Computer Linguistics.