Natural Language Engineering **1** (1): 1–23. Printed in the United Kingdom

Possession Identification in Text

CARMEN BANEA and RADA MIHALCEA
Electrical Engineering and Computer Science,
University of Michigan, Ann Arbor, MI, US.
e-mail: carmennb@umich.edu, mihalcea@umich.edu

(*Received 10 May 2017; revised 2 February 2018; accepted 5 February 2018*)

Abstract

Just as industrialization matured from mass production to customization and personalization, so has the Web migrated from generic content to public disclosures of one's most intimately held thoughts, opinions and beliefs. This relatively new type of data is able to represent finer and more narrowly defined demographic slices. If until now researchers have primarily focused on leveraging personalized content to identify latent information such as gender, nationality, location, or age, this article seeks to establish a structured way of extracting possessions, or items that people own or are entitled to, as a way to ultimately provide insights into people's behaviors and characteristics. We introduce the new task of "possession identification in text," and release a novel dataset where possessions are marked at different confidence levels. We present experiments and results obtained when seeking to automatically identify and extract possessions from text.

Keywords: possession identification, latent attribute extraction, author profiling

1 Introduction

With the introduction and adoption of Web 2.0, the Internet has become a forum where users voice their opinions and feelings through comments, reviews, blogs, microblogs, status updates, and other forms of online participation. This growing and diverse unstructured stream of information blends for the first time consumer demographics, with lifestyle information and choices, user opinions, as well as mentions of items that are owned or liked by the user.

Our research is motivated by the affective-cognitive consistency model (Rosenberg, 1956; Rosenberg, 1968), a branch of cognitive consistency theory that not only hypothesizes that people are motivated to seek a coherent state both internally (at the level of thoughts, beliefs, feelings, and values) and externally (through attitudes and behaviors), but also that individuals gain more motivation in achieving a consistent state so that others perceive them to be consistent. This particular model implies that in a public setting, where others are reading a person's online content,

† Corresponding author.

the blended data forms not only a raw user profile, but it encodes a coherent user signature, which, if mined successfully, can carve out very narrowly defined clusters of individuals who live, think and act alike. An empirical study conducted by (Stecher & Counts, 2008), showing that readers are significantly better able to infer and recall latent personality dimensions embedded in an online profile, compared to the scenario of an artificially altered profile, further demonstrates the ability to extrapolate a coherent author signature from user content.

Motivated by the potential of ultimately analyzing and evaluating people's behaviors and characteristics as they relate to their ownership of a given object, this paper lays the groundwork of *identifying such owned objects*. To this end, we propose the concept of "possessions," or textual representation of items that somebody owns or is entitled to, which we define in detail in Section 3.1. In terms of ethical concerns in regard to possession extraction, we believe they are in line with those entailed by the more general task of information extraction, as we are only highlighting possessions that are already mentioned in the text, and we do not uncover hidden traits. Nonetheless, if this information will be leveraged and joined with latent attribute extraction such as gender, occupation, age, etc. to analyze people's behaviors and characteristics, it should be employed only to generate aggregate models, where a group of users are represented as a whole and not as unique individuals, in order to protect their privacy.

Our work aims to answer two research questions. First, is the task of possession identification well defined, with a set of labeling guidelines that can enable the consistent annotation of possessions in text? We first define possessions, and introduce extensive guidelines for the annotation of possessions in text. From a set of blog data, we identify approximately 800 possessions, conduct inter-annotator agreement analyses, and construct an annotated dataset that we release to the research community to facilitate further work on this task. Second, can we build a machine learning framework that can automatically identify such possessions in text? We proceed by modeling the possession identification task as a supervised learning problem. To do so, we identify a set of features and present an extensive analysis performed against three different baselines, over which our results show significant improvements. Using our best performing model, we further conduct a pilot study that looks at the correlation between people's gender and their possessions.

This article starts out with an overview of the relevant literature in Section 2. Section 3 then defines possessions and highlights a set of considerations needed to ground the possession annotation process, ultimately resulting in a proposed annotation schema. We introduce the dataset we collected and the various statistics that pertain to it in Section 4. In Section 5 we focus on modeling the task as a supervised learning problem, by starting with feature engineering, introducing the experimental setup, and presenting our results and discussing our findings. We apply our possession extraction system in a pilot study (see Section 6) that seeks to identify correlations between items that people own and their gender. Ultimately, Section 7 concludes by summarizing our findings and pinpointing future work directions.

2 Related Work

To date, research focusing on extracting latent user-descriptive attributes from microblogs has been mostly centered around Twitter, as it is a service with a high adoption rate, where many of the users share their tweets publicly. Some of the attributes targeted for extraction are demographics related, such as gender and age (Burger & Henderson, 2006; Mukherjee & Liu, 2010; Rao *et al.*, 2010; Pennacchiotti & Popescu, 2011; Burger *et al.*, 2011; Van Durme, 2012), political affiliation (**?**; Rao *et al.*, 2010; Pennacchiotti & Popescu, 2011; Zamal *et al.*, 2012; Cohen & Ruths, 2013), affinity with a given business entity (Pennacchiotti & Popescu, 2011), mental health, intelligence, relationship status, religion (Volkova & Bachrach, 2015; Volkova & Bachrach, 2016), lifestyle (Hu *et al.*, 2017), and more. While at the beginning, research was primarily conducted over literary text, around 2010, publications started focusing on Twitter, primarily because in addition to the text of the tweets, it also gives access to information about the user from the self-authored user profile, and most importantly because it allows access to the social network of a given user: namely friends (people the user follows), mentions (people the user mentions in the tweets by their handle "@username"), and followers. As such, the machine learning algorithms employed have learned over a mix of features extracted from (1) the microblog text itself such as unigrams, bigrams, trigrams, hashtags (Cheng *et al.*, 2010; Pennacchiotti & Popescu, 2011; Zamal *et al.*, 2012; Volkova & Bachrach, 2015), (2) from the short bio / profile information (Pennacchiotti & Popescu, 2011; Zamal *et al.*, 2012), (3) from a user's behavior in the social media platform: tweet frequency, number of retweets, number of replies (Pennacchiotti & Popescu, 2011; Volkova & Bachrach, 2015), and (4) the underlying social graph of friends and followers (Zamal *et al.*, 2012; Volkova & Bachrach, 2016). The social graph has been modeled through its ability *to encode topography* – showing who are the producers of information (those who tweet often and have a wide following) and consumers of information (those who tweet rarely, and have few followers), and how based on different heuristics, nodes within a network may be predictive of various latent attributes of the user in question (Zamal *et al.*, 2012) –, as well as through its ability *to act as an oracle*, allowing words, hashtags, or topics that are more descriptive of a given subgroup to emerge and act as differentiating features (Pennacchiotti & Popescu, 2011; Zamal *et al.*, 2012).

A major impediment in conducting research in this area has been the lack of training data. Ground-truth labeling has been leveraged from self specified information in a user profile, such as location (Cheng *et al.*, 2010), gender and ethnicity (Pennacchiotti & Popescu, 2011), but this was shown to be an unreliable venue, as (Cheng *et al.*, 2010) report that only 26 percent of the users specify their precise location (city), while the remaining provide a much wider area (state, country), as well as imaginary places (i.e. Wonderland), and (Pennacchiotti & Popescu, 2011) disclose that they were only able to identify the gender of 80 percent of the users, and that with a low accuracy, while the ethnicity could be inferred for only 0.1 percent of the users. Researchers have also employed crowd-sourcing services such as Mechanical Turk, where human judges would decide on these attributes by reading a user's profile/blog (Rao *et al.*, 2010; Burger *et al.*, 2011), from their profile picture (Pennacchiotti & Popescu, 2011; Ciot *et al.*, 2013; Liu & Ruths, 2013), or username (Zamal *et al.*, 2012). More recently, researchers have explored the option of using distant supervision by linking information from external knowledge sources such as Google+ or Facebook to existing Twitter profiles, and thus achieved ground truth labeling for attributes such as spouse, education and job that are not supported by the Twitter platform and involved no human annotation effort (Li *et al.*, 2014). In addition, (Hu *et al.*, 2017) have employed automatic posts to Twitter made by applications such as Foursquare to obtain precise user location based on points-of-interest (restaurants, museums, etc.). As a complementary method (Volkova & Bachrach, 2016; Volkova & Bachrach, 2015) have employed predictions made by machine learning algorithms trained on a smaller set of data enhanced with socio-demographic user attributes identified through a crowd-sourcing annotation task, to generate a larger training set, which after subsequent training, yields results that are seemingly better or with a similar performance to that of single stage learning algorithms. While such an approach has not been evaluated for its predictive accuracy in regards to gold standard socio-demographics characteristics pertaining to a given user, the evaluations at an aggregate level attain state-of-the-art results. In addition, online services such as the `genderize.io` API have emerged and provide a gender probability distribution for a social media username.

Work to date has mainly focused on attributes that have a narrow range of values, typically binary. For example, (Rao *et al.*, 2010; Pennacchiotti & Popescu, 2011; Zamal *et al.*, 2012) have looked at more generic user characteristics such as gender (male / female), age (young adult / mature), political affiliation (democrats / republicans), but other latent user attributes have been explored as well, such as ethnicity (African-American vs. all other ethnicities) (Pennacchiotti & Popescu, 2011), support for a given business (Starbucks fans vs. everyone else) (Pennacchiotti & Popescu, 2011), regional origin (North or South Indian) (Rao *et al.*, 2010), etc. Volkova and Bachrach (Volkova & Bachrach, 2015; Volkova & Bachrach, 2016) go even further, seeking to predict a dozen socio-demographic user descriptors, yet each of these are also represented in a binary form (i.e. with children / without children, in a relationship / single, Christian / unaffiliated, African American/ Caucasian, below & average intelligence / above average intelligence, satisfied/dissatisfied with life, etc.). Such choices have in a way simplified the problem, as the random baseline becomes 50 percent, while also not truly allowing for a meaningful attribute extraction, i.e. being able to determine that a user is Buddhist, of East Asian ethnicity and in the 40s. A notable exception has been the prediction of location based on user content and social network, which is able to model from coarse city-level granularity (Cheng *et al.*, 2010) all the way to within 100 meters of the actual location of a target user in 20 minutes increments (Sadilek *et al.*, 2012). In contrast, our work focuses on extracting a novel attribute type (possessions) that has an open vocabulary, since any concrete noun can be a possession. To our knowledge, this is the first study that seeks to identify object ownership.

3 Possessions

3.1 What Are Possessions?

We define possessions as textual representations of physical, concrete objects that could be considered to be someone's property such as electronics, clothes, furniture, etc., or of items to which somebody is entitled to due to his / her position or social standing, such as an employee to his cubicle, or a king to his throne. Possessions, however, cannot be human beings, as people can exercise free will: "my mother" appearing in a given context does not render the denoted person a possession, despite the preceeding possessive article.

3.2 Possession Annotation

Marking possessions consistently in text requires establishing a thorough set of guidelines that encompass several considerations such as:

Ownership. We identify possessions with respect to the author of the utterance. For example "I left my laptop in the car," suggests that the writer owns a laptop; however, the same context sheds no light on whether the car is implicitly his / hers as well; as such, considering the limited information, the automobile is not considered a possession. Another aspect to examine is the fact that a possession can exhibit joint ownership. For example, in the sentence "My husband and I own a beautiful house," the house is an object to which both parties are entitled, thus the object is a possession of the speaker, as well.

Time frame. A given object needs to be possessed by the writer at the time of the utterance in order to be considered a possession. Items owned in the past or whose current status is unknown are not considered possessions. This time frame consideration also allows accounting for negations, irrealis and sarcastic statements in text. For example statements such as "I never had a car" (negation), "I always wished I had a car" (irrealis), "Of course, I have a personal chopper!" (sarcasm), can be accurately processed as containing no possessions.

Identifiability. When annotating possessions, one of the main aspects to consider is the identifiability of the expressions being annotated. Let us consider the following sentence: "I left my shoulder bag in the car." The identified item cannot be simply "bag," as that would be too ambiguous; are we talking about a backpack, a shopping bag, a beauty bag, purse, or luggage? For this reason, the shortest span we can annotate, which will also provide a precise idea of the actual item being possessed, is "shoulder bag." In addition, generic words such as "things," "items," "collections" should not be annotated, as they are not identifiable.

Document level consistency. Furthermore, the identification of possessions is considered with respect to the entire document. Items whose ownership status may have been unknown in the initial passages of a document may be attributable to the writer once all the context has been taken into consideration. As such, all mentions of those items in the document are properly resolved upon a second pass. Similarly, items that change ownership within the document are properly resolved based on the last known owner of the item.

Concrete nouns. Possessions are concrete nouns representing objects that can occupy a physical space. Another way to think of such objects is by considering their picturability potential. Nouns such as cup, fork, desk, computer, are concrete, and therefore potential possessions, while nouns such as love, happiness, goals, etc. are not. Another consideration is that even if an item exists only in a virtual world, such as an email, blog, document or photograph, the fact that such items are printable and therefore can become tangible, renders them potential possessions.

Resolution scope. While items are decomposable in terms of the constituent parts, we consider possessions with respect to the whole. A cellphone may contain a screen, case, etc., but the possession being identified is the cellphone. Similarly, body parts are not annotated because they resolve to a person, and as mentioned earlier, persons are not items in order to be considered property.

These considerations, together with the feedback received from the annotators were incorporated into an extensive set of guidelines that we are releasing with the article[1].

3.3 Annotation Format

All possessions that meet the considerations mentioned earlier are marked in the text using an XML schema. Each possession span is enclosed within an $< object >$ tag, which can take several attributes:

- *id*: a unique number identifying the possession within the document. It starts at 0, and is incremented every time a new possession is identified within the same document. Multiple mentions of the same possession within a document are resolved to the same id.
- *value*: an expression describing the item type as found explicitly or implied from the text. Partial textual references to an item are resolved and cross-referenced to an identifiable object (as required by the identifiability constraint). All items having the same *id* also have the same *value*.
- *type*: "perm" (permanent) / "temp" (temporary); refers to how persistent the possession is. If a possession lasts less than one day, we consider it "temporary," otherwise, it is perceived to be "permanent." For example, if the possession is a perishable item (ice-cream, coffee) or an item that is not expected to last (ice), the type is set to "temporary."

Let us look at the following example:

" I left my green shoulder bag in the car."

Once annotated, the sentence becomes:

[1] See "Possession identification guidelines" at `http://lit.eecs.umich.edu/research/downloads/`

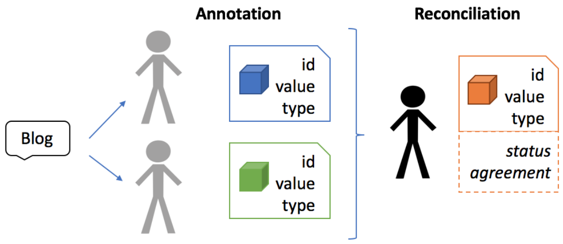

Fig. 1. Two level annotation process. The last judge reconciles the annotations and ads the final status and the agreement attributes.

I left my <object value = "shoulder bag" id = "0" type ="perm"> green shoulder bag </object> in the car.

The *span* of the annotation is "green shoulder bag," as it unequivocally establishes the object that belongs to the writer. However, the *value* of the object is devoid of personalized information, thus allowing cross owner profile analysis (both A and B may possess a shoulder bag, but only A's bag is green). The *id* of the item is 0, assuming that it is the first possession encountered in the blog, and its *type* is permanent, as it is an object that will last the owner an extended period of time.

As the data is initially annotated by two judges, and then reconciled by a third judge, at the end of the reconciliation stage two additional attributes are introduced:

- *status*: denotes the fact that the third annotator casts a confidence vote for a particular marked span by adding a *status* = "*final*" attribute.
- *agreement*: lists the first name initial of the particular annotators who identified an object. For example, for *agreement* ="*em*", two judges concur in their determination.

4 Dataset

Using the annotation guidelines described above, we annotated a set of 799 possessions that appeared in 27 blog posts authored by different bloggers encompassing 120 thousand tokens, with an average length of approximately 4,500 tokens per blog. These were collected from the Web between May and July 2015, and pertain to categories such as lifestyle, travel, health, weddings, shopping, as shown in Table 3. Unsurprisingly, due to the personal nature of the blogs, most possessions are associated with the lifestyle category.

Three senior students majoring in linguistics from the University of Michigan participated in the annotation. The set of blogs was split into three parts, and through rotation, each subset of 9 blogs was assigned to a different annotator to reconcile, upon receiving individual annotations encoded by the other two judges. The third annotator was the only one who could specify the *status* and the *agreement* attributes in the $< object >$ tag (see Figure 1). Out of a total of 799 annotations, the first two annotators agreed on 354 annotations (44.4 percent). A match was considered when the spans for a possession overlapped. Upon reconciliation by the third annotator (who decided on the length of the span to be included in the final annotation document, as well as the proper values for its attributes), 615 annotations received *status* "final" (77.1 percent), indicating that upon considering the views of the other two annotators, the third annotator decided to support the annotation (marking it as "final"), or to reject it (by not changing its status to "final"). We also encountered 9 cases where the third annotator missed updating the status to "final", despite agreeing with the annotation, as marked by his/her first name initial in the agreement field. Upon correction, 624 annotations received a final status, representing 78 percent of the possessions marked.

Several reasons why the annotators disagreed on marking a possession span were identified: (1) the author's statement was perceived as being *generic*, and therefore the possessions mentioned were considered not indicative of actual possessions of the author, (2) the author was recounting events that took place in the past, and there was not enough information available to know the *current status* of the possession, (3) an annotator incorrectly attributed an item to the author. Table 1 exemplifies the above scenarios.

The dataset annotation can be utilized from two perspectives. (1) As the result of a *pipeline* or a two step process, where two annotators mark the text first, and then a third reconciles their markings; in this case the annotation receives the *status = "final"* attribute, as it shows that the final judge agreed with the annotations. In this scenario, we can consider that those annotations whose status is not updated are weaker, as the third annotator does not cast a confidence vote in their support. (2) The annotations can also be seen as the result of a *single stage democratic process*, where annotators have equal power in casting a vote toward considering a span as a possession. In this later view, the *agreement* attribute, which specifies the first name initial of each of the annotators that agreed on marking the span can be used to derive gold, silver and bronze-standard annotations. Based on the reconciled annotations, we are releasing the dataset with several confidence levels: the gold-standard, where all three annotators agreed on the possession being marked (345 possessions / 43.23 percent), the silver-standard, where at least two annotators agreed (583 possessions / 73.06 percent), and the bronze-standard, consisting of all the annotations made (799 possessions / 100 percent).[2]

Table 2 lists the top ten possessions identified during the annotation study. The value of the possession (the noun entry in the first column in Table 2) was selected by the annotators based on the document where the possession occurred, and the annotation span represents various mentions of the possession in text. Since these mentions were all reconciled at the document level, the same mentions may trigger different possessions, such as "house" and "home" triggering *house* in one scenario, while "home" triggers *home* in another. We should note that some of the possessions are implied through the use of verbs, such as "driving" implying the possession of a car, or "called" / "phoned", implying the possession of a phone. Overall, we encountered an average of 29.56 possessions per blog; these are mostly expressed

[2] The dataset is publicly available from `http://lit.eecs.umich.edu/research/downloads/`

Example	Annotation Discrepancy Reason
Getting this stuff right is so hard. Especially when what is right changes every day. One day you are to push and encourage and demand that they stand out from the pack. Be the best! The next day you are to be accepting and allow the child to establish his own personality and identity. You are to study with them, and teach them good studying habits. Oh no! That was last week. These children are old enough to be responsible for their own lessons. Studying with them will make them unable to study on their own. Just take the video games away so they wont have violent tendencies. Great! Now they dont have any friends and cant pick up MMs because their hand/eye coordination is so horrible.	Statement perceived as generic and the possessions mentioned were considered not indicative of actual possessions of the author.
making ricotta, which was something I had wanted to do for a very long time ever since I saw the recipe in Saveur six or so years ago and ripped out the pages	Status of possession unknown at present time.
The boys put on their skates and watched the neighbor boy and it was over.	Item incorrectly attributed to be the ownership of the writer.

Table 1. *Sample excerpts from blogs showcasing the reasons behind discrepancies in annotations. The items identified as possessions by the annotator are underlined.*

through nouns, at an average of 1.07 nouns per span. 89 instances where the possession span included a verb were identified, representing 11.15 percent of the total number of possessions.

Out of the 799 marked possessions, only 199 (25 percent) of them are precluded by possessive determiners (my, mine, our, ours) in a window of three words preceeding the possession annotation, while the entire dataset contains 1183 possessive determiners. This shows that only 16 percent of the possessive determiners are followed by a possession annotation, signaling the difficulty of the task. Furthermore, while the dataset contains 2,645 occurrences of personal pronouns (I, we), only 82 of these appear in a window of three words prior to a possession span.

Some of the most often encountered possessions in the gold-standard are: house (27), prosthetic leg (16), phone (16), wedding dress (15), and car (11). For the

Possession	Frequency	Annotation Span
house	13	house, home, place
blog	10	foodie blog, blogging, post, blog, blog post
photo	8	photo, picture
car	8	driving, car, drive
bed	7	bed
home	7	home, at-home
shoes	6	shoes, Toms
clothes	6	clothes, regular clothes, got all dressed up, outfit, running clothes
phone	6	speakerphone, called, phone, phoned

Table 2. *Top 10 most frequent possessions encountered in the dataset.*

Category	# of blogs	# objects	Average
Lifestyle	8	265	33.12
Travel	4	104	26.00
Other	3	52	17.33
Health	3	105	35.00
Wedding	2	69	34.50
DIY	1	32	32.00
Real Estate	1	6	6.00
Parenting	1	14	14.00
Pets	1	23	23.00
Fitness	1	38	38.00
Shopping	1	6	6.00
Medical	1	84	84.00
Overall	27	799	28.44

Table 3. *Distribution of blogs over categories. Columns* **1***: Category;* **2***: Number of blogs;* **3***: Number of possessions identified in a category;* **4***: Average number of possessions per blog in a given category.*

silver-standard, additional high-frequency possessions are items such as photo (18), blog (16), and glucose monitor (10), while for the bronze-standard we have loaner car (10), gym (10) and picture (10). We did notice that over-specialized blogs, such as those focused on medical experiences or weddings, have a very narrow and frequently used possession vocabulary, as the authors seem to maintain a log of the activities for their own use, and not necessarily for the enjoyment of the reader.

Out of the 11,475 nouns that occur in our dataset, identified using the automatic part-of-speech tagger from the Stanford CoreNLP package (Manning *et al.*, 2014), only 694 of them were found within spans marked as possessions by the annotators, representing a proportion of 6.05 percent. This implies that despite the personal nature of blogs, relatively very few nouns actually represent author possessions, further indicating that the task of automatically extracting possessions from text is quite challenging.

5 Automatically Extracting Possessions from Text

We frame possession identification as a two-class machine learning task. Since approximately 90 percent of the possession spans include automatically identified nouns, we construct instances at the noun level; the nouns appearing within a possession span are considered possessions, while the remaining nouns receive a non-possession label.

Using the Stanford CoreNLP library, a blog is split into sentences, which are then tokenized, part-of-speech tagged and parsed using a dependency parser. For every identified noun from the set of 11,475 nouns labeled by the Stanford CoreNLP tool, we derive a set of features as described in the following section.

5.1 Features

The following features are generated for each noun in the data set:

Context words. Within the same sentence, a span of five tokens to the left and right of the target noun is used to extract word unigrams, bigrams and trigrams.

Context POS. Same method as above, yet instead of considering the words themselves, we retain their part-of-speech tag, and construct POS unigrams, bigrams and trigrams.

Semantic dependency parsing relations. We obtain all the dependency parse relations in which the target noun participates, by finding all the incoming and outgoing edges to other tokens in the same sentence. These are encoded as features using the relation type, the edge direction, and the related token.

Let us consider the following example:

Seeing something so far out of what my <u>surroundings</u> taught me has shocked me.

The underlined words represent tokens identified as nouns by the automatic POS tagging system. To illustrate the attributes we extract, we will consider the noun "*surroundings*," as it takes the role of either dependent or governor in the dependency relations in which it participates, namely: (1) (surroundings/NNS, my/PRP$) nmod:poss and (2) (taught/VBD, surroundings/NNS) nsubj. In the case of the first relation, we have one outgoing edge, from the possessive pronoun *my*, whose relationship to *surroundings* is that of a possessive nominal modifier. In a lexicalized form, this relation is encoded as "nmod:poss-my/PRP$", where "-" represents that the edge was outgoing. For the second relation, we have an incoming edge, from the verb *to teach* identifying *surroundings* as a nominal subject. This relation becomes "nsubj+teach/VBD", where "+" signifies that the edge was incoming. By using the dependency relation and its directionality, as well as the lemmatized version of the paired token (to increase coverage) and its part-of-speech, we are able to formulate precise scenarios where a given object may constitute a possession. In order to further generalize our dependency rules, we also generate unlexicalized features, which consist of the same elements as their lexicalized counterpart, except the paired token; the unlexicalized form becomes: "nmod:poss-/PRP$", and "nsubj+/VBD", respectively.

Preceding possessive marker. This is a binary feature, indicating if a possessive marker associated with the first person singular or plural appears in a five word span before the target noun. We only look at first person forms since the possessions in our dataset are annotated with respect to the writer of the blogs. Possessive markers may be either possessive determiners (my, our) or possessive pronouns (mine, ours).

Nearby adjective. When writers mention a possession in text, such as a new blouse or cellphone they purchased, they often times use adjectives to describe the item in more depth. In addition, adjectives are words that modify or describe nouns. Our assumption is that the more concrete a noun is, the more it enables the presence of adjectives that describe it. Concrete nouns are sensory-oriented since they occupy a physical space. Since possessions are physical objects, then the presence of adjectives may be considered a marker for potential possessions. As such, we use the binary feature reflecting if an adjective is identified in the vicinity of a target noun (in a span of five words left and right), aiming to model whether the noun is more likely a possession.

Category information. This attribute seeks to capture whether a noun may denote a concrete object, and thus potentially qualify as a possession. Category information is retrieved by querying Walmart's online product database.[3] Two attributes are extracted, one pertaining to the main product category, and one retaining the subcategory of the product in question. For example, if we consider the noun "iPad", the main product category will be *electronics*, while the full path will be *electronics - iPad & tablets*. Despite our expectation that using a product database would filter out abstract or proper nouns, this was not the case, as likely categories were returned for many entries; for example, querying for "patience", returned *Party & Occasions - Christmas Decor* or the proper noun "David" triggered the *Food - Snacks, Cookies & Chips* category.

Concreteness score. Another way we try to encode whether a noun may be concrete is by using a concreteness score. We use the Free Association Norms database (Nelson *et al.*, 2004), consisting of 5,019 cue words for which 6,000 participants provided a free text entry of the first word that came in their mind that was related or associated with the stimulus word. 3,278 of these words were annotated with a concreteness score ranging from 1 (extremely abstract) to 7 (fully perceptible with the senses), and out of these, 2,305 were nouns. If our target noun appears among this latter group, we include its concreteness score as an attribute.

NER information. To further distinguish between nouns that may represent a

[3] `https://developer.walmartlabs.com/`

person, organization, location, etc., we also include a feature that encodes whether the target noun belongs to a particular named entity category. We used the Stanford CoreNLP named entity recognition module, which is able to identify thirteen such categories: named entities (person, location, organization or miscellaneous), numerical mentions (money, number, ordinal, percent) and temporal mentions (date, time, weekday, duration, set).

Levin verb classes. To model the functional similarity of the verbs surrounding the target nouns, and therefore their ability of carrying arguments that may be possessions, we include features based on the verb taxonomy proposed by (Levin, 1993). Levin analyzed and classified English verbs from two perspectives: verb alternations (8 major groups) and verb classes (41 major groups), each consisting of further subcategories. Each verb appears at least once in the alternations set and once in the classes set. For alternations, she looked at the capacity of verbs to entertain a variety of object alternations, such as for example the *locative alteration of "putting" subtype*. Levin (Levin, 2006) exemplifies it as follows:
"Jill sprayed paint on the wall.
Jill sprayed the wall with paint."
Here we can see how the verb *sprayed* implying putting something on a surface, allows the arguments *paint* and *wall* to have a different position yet to convey the same exact meaning. Similarly, for classes, she grouped verbs based on their event structure.

To exemplify, Levin (Levin, 2006) considers verbs of removal once looking through the lense of the manner in which the action was accomplished: *sweep*, *wipe*, and once looking at the result of the action: *clear*, *empty*.

For a given verb in the vicinity of a target noun, we query the taxonomy and extract four features: two for the alternations (one coarse grained and one fine grained), and two for the verb classes (one coarse grained and one fine grained). The coarse grained features consist of only the index of the major verb group, while the fine grained ones retain the entire subcategorization path. In addition, these features are extracted based on the position of the verb, namely before or after the target noun. In total, eight Levin-based verb features are extracted.

For example, for the verb *drive*, we extract the alternation coarse grained grouping (1) and the fine grained grouping (1.1.2.2) placing it in the *induced action alternation* set, together with verbs such as *canter*, *fly*, *gallop*, etc. To exemplify the idea behind this feature choice, *canter* and *gallop* could point to the possession of a horse, while *fly* could imply the possession of a plane. In terms of class grouping, we encounter the verb in the taxonomy at position 11.5 paired with other verbs such as *barge*, *bus* and *cart*. We extract the class coarse grained grouping as 11, and the class fine grained grouping as 11.5. This allows us to create a common feature representation for the verbs of type "drive", allowing us to potentially identify implied possessions such as *barge*, *bus* or *cart*.

Verb tense. Our guidelines specifically instruct the annotators to mark possessions that the writer of the blog owns or is entitled to as of the moment of the utterance. In order to capture this information as a feature, we look at the part-of-speech annotation, and identify as “present” only the verbs tagged as VBG (verb gerund / present participle) or VBP (verb non third person singular present), while all others are defaulted to “other.” We keep track separately of the tense occurring before and after the target noun.

5.2 Experiments

We split the data into 27 folds corresponding to each of the 27 annotated blog posts, to ensure that possessions from the same blog are not split between training and test. This decision is motivated by the fact that in the span of a single blog, there are often times multiple mentions of a given possession; by considering an entire blog either in test or in the train set, but not split across the two, we ensure that no instance in the train data will match one in test, thereby potentially making it artificially easier for the test instance to be labeled. Furthermore, 27 folds allow us to use the maximum amount of data for training, and therefore use the highest number of possession annotations to learn from.

We experimented with several machine learning algorithms:[4] support vector machines (Platt, 1999), K-nearest neighbors (Aha *et al.*, 1991), decision trees (C4.5 (Quinlan, 1993)), feed-forward neural nets (Hornik, 1991) and Naive Bayes. In order to avoid overfitting the data, these algorithms were evaluated only on the first two test folds (out of the 27); all the algorithms except for the Naive Bayes either made no assignment of test instances to the possession class, or they did so very sparingly.[5] Consequently, the results reported in this section are based on a 27 fold cross-validation using Naive Bayes, and are presented in Table 5. Throughout the rest of the paper, we will reference a particular entry in this table by using a parenthesized line index key.

5.3 Evaluation Metrics

As typical for this type of evaluation scenarios, we will employ per class precision, recall and F-measure. Let us consider a binary classification truth table (see Table 4). A true positive case is then when a possession is correctly labeled as a possession by our algorithm, a false positive is when a non-possession is incorrectly labeled as a possession by the algorithm, a true negative is when a non-possession is correctly identified as non-possession, and a false negative is when a possession is incorrectly labeled as a non-possession by the algorithm. Since the possession extraction algorithm does not extract a span at this time (such as “green bag”), but rather only the noun (“bag”), we did not have to account for span alignment considerations for now.

Precision (P) represents the fraction of correctly identified instances (true

[4] Implemented in the Weka machine learning library (Hall *et al.*, 2009).
[5] Resampling the data with a bias toward uniform class distribution did not achieve a positive impact.

		Actual class	
		Possession	Non-possession
Predicted class	Possession	TP	FP
	Non-possession	FN	TN

Table 4. *Binary classification truth table.*

positives[6]) to the total number of instances labeled by the system (true positives and false positives), i.e.:
$P = \frac{TP}{(TP+FP)}$.

Recall (R) is the fraction of correctly identified instances (true positives) to the total number of instances that were relevant (true positive and false negatives), i.e.:
$R = \frac{TP}{(TP+FN)}$.

F-measure (F) is the harmonic mean of precision and recall, allowing to capture the duality of their behavior (typically, as precision increases, recall decreases, and the reverse) in a snapshot, i.e.:
$F = \frac{2*P*R}{(P+R)}$.

5.4 Baselines

We compare the results of our system against three baselines.

Unsupervised baseline (1). The unsupervised baseline uses a 50:50 split assignment of the two classes to make a prediction, and the labels are randomly assigned to each instance. As such, it is able to achieve an F-measure of 11.5 percent, yet its accuracy is about 50 percent.

Supervised baseline (2). The supervised baseline uses information from the learned class distribution, and labels all test samples with the majority class prediction. While its accuracy is the highest, at 94 percent, its possession class F-measure is 0 percent.

Possessive marker baseline (3). This baseline always assigns nouns that are preceeded by a possessive marker to the possession class. From the 11,475 automatically identified nouns, 694 of them meet this criteria. Such rule-based assignment is able to achieve a possession class precision of 11.4 percent, with a recall of 28.2 percent, and an F-measure of 16.2 percent.

[6] True positive, false positive, true negative, false negative are abbreviated as TP, FP, TN, FN, respectively.

Variation	Possession			Non-possession			Overall
	P	R	F	P	R	F	Acc
(1) unsupervised baseline	6.4	53.6	11.5	94.3	49.7	65.1	49.9
(2) supervised baseline	0.0	0.0	0.0	94.0	100.0	96.9	94.0
(3) possessive marker baseline	11.4	28.2	16.2	94.9	85.8	90.1	82.4
(4) text	13.6	8.9	10.8	94.3	96.4	95.3	91.1
(5) pos bigram	8.7	**18.4**	11.9	94.3	87.6	90.8	83.4
(6) pos trigram	10.3	13.3	11.6	94.3	92.6	93.5	87.8
(7) dep lex	**14.4**	10.5	**12.2**	94.3	96.0	95.2	90.8
(8) dep unlex	11.7	9.1	10.2	94.2	95.6	94.9	90.3
(9) low Levin prec. v. alt. class	8.3	11.7	9.7	94.2	91.7	92.9	86.9
(10) dep (un)lex	**13.2**	23.3	16.8	94.8	90.1	92.4	86.0
(11) text + pos unigram**	12.9	13.0	13.0	94.4	94.4	94.4	89.5
(12) text + pos bigram**	11.0	**30.0**	16.1	94.9	84.4	89.4	81.1
(13) text + pos trigram**	12.7	25.7	**17.0**	94.9	88.7	91.7	84.9
(14) text + dep (un)lex**	14.5	32.7	20.1	95.3	87.6	91.3	84.2
(15) + pos best (trigram)**	12.9	42.9	19.8	95.7	81.3	87.9	79.0
(16) + main category**	15.5	33.4	21.2	95.4	88.3	91.7	85.0
(17) + sub category**	17.7	37.5	24.0	95.7	88.8	92.1	85.7
(18) + concreteness	17.2	38.2	23.7	95.7	88.2	91.8	85.1
(19) + main & sub category**	19.2	40.4	26.1	95.9	89.1	92.4	86.1
(20) + concreteness	**22.4**	**46.7**	**30.3**	96.3	89.6	92.8	87.0
(21) dep (un)lex + main/sub cat	19.3	34.2	24.6	95.5	90.8	93.1	87.3
(22) + concreteness	**24.5**	44.1	**31.5**	96.2	91.3	93.7	**88.4**

Table 5. *Evaluation results, including per class micro precision (P), recall (R) and F-measure (F), and overall accuracy (Acc). The numbers in parentheses are used to cross-reference a particular variation in the text. The marks ** denote statistically significant results (*$P \leq 0.0001$*) when compared to the variation that does not include the last term. The numbers in bold represent the highest value attained in a given block.*

5.5 Results & Discussions

Table 5 provides an overview of our results. While the three baselines we propose are able to reach an F-measure of at most 16.2 percent, ultimately, using our feature set results in an almost two fold increase in F-measure, to 31.5 percent (22).

From all the features we extract, few of them are able to be used in a stand-alone scenario to classify possessions.[7] The ones that have a possession class F-measure of approximately 10 percent are: context features (4), part-of-speech bigram and

[7] Since non-possessions are the majority class, and defaulting to this label results in an

trigram (5, 6), lexicalized and unlexicalized dependencies (7, 8), and a Levin verb taxonomy-based feature (9). The highest possession class precision with an acceptable recall is 14.4 percent, pertaining to the lexicalized dependencies (7). With this information as a starting point, we further evaluate whether pairing high-performing features translates into better modeling of the classification task. One step is to check whether lexicalized and unlexicalized dependencies capture similar or orthogonal information; if each one individually has a precision for the possession class of 14.4 percent (7) and 11.7 percent (8) respectively, and a recall for the possession class of approximately 10 percent, together, they are able to retain their precision (at 13.2 percent), while more than doubling their recall (23.3 percent) (10). Further pairing them with any part-of-speech based features (whether unigrams, bigrams, or trigrams), always results in a decrease in precision, seemingly signaling that dependencies are already capturing much of the information needed from the functional aspect of the contextual words. Similarly, grouping context-based features with part-of-speech (whether unigram, bigram or trigram) results in a drop in precision for the possession class (11, 12, 13). A meaningful threshold jump occurs by pairing contextual and dependency-based signals (14), where precision for the possession class becomes 14.5 percent, surpassing the precision of both the component parts, and recall moving outward significantly, to 32.7 percent. Adding typical untargetted text processing features to the context unigrams - dependency parsing mix always decreases performance (15), signaling that an upper bound with the information that this type of features can capture has been reached.

Since the main trait of possessions is that they are objects that can occupy a physical space, it makes sense that seeking to model that information either through a product database or a concreteness score as provided by a lexical resource may prove helpful. Adding information modeling the former (19), increases possession class precision to 19.2 percent, while also covering more cases, with a 40.4 percent recall, while adding just the concreteness score (18) improves precision to 17.2 percent, and recall to 38.2 percent. Allowing both these feature types to work together (20), results in the best possession classification scenario, with a possession class precision of 22.4 percent and a recall of 46.7 percent (F-measure of 30.3 percent). At this point, it is not clear whether the contextual features are still needed, as without them (21), the model is able to achieve a possession class precision of 24.5 percent, while experiencing a small drop in recall, to 44.1 percent (F-measure 31.5 percent), a 93.7 percent F-measure for the non-possession class, and 88.4 percent in the overall accuracy; all this with less than half the number of features (8,429 features including context unigrams versus 3,632 features without).

The annotation guidelines direct the analyst to only mark possessions that the writer owns as of the time of the utterance. Interestingly enough, infusing information into the model that seeks to cover verb tense does not have an impact on the classification task. We assume this is the case because we are exploring the blog-genre, where writers focus on current events, and therefore most of the possessions

accuracy of over 90 percent, the discussion that follows will focus on achieving better results on the possession class.

they mention are anchored in the present. Furthermore, using the Levin derived taxonomy to group verbs that experience a functional similarity does not achieve a positive impact on the classification task.

It is surprising that possessive markers (such as the possessive determiner and pronoun) are not among the reliable features in detecting a possession, yet in retrospect, it makes sense that their usage accompanying both abstract and concrete nouns is too undiscerning. For example, “my life” or “my thoughts” are never marked as possessions by the annotators, as they are abstract concepts. Similarly, the nearby adjective feature is also too generic to capture information pertaining to possessions.

The named entity recognition feature minimally impacts the best performing feature set (22), achieving an increase in the possession class recall by more than 1 percent (to 45.2 percent), yet exhibiting a small drop in precision by 0.5 percent (to 24 percent). From the data we have, it remains unclear whether adding this feature to the mix is useful.

5.6 Error Analysis

In order to gain a deeper insight in the differences between the actual and predicted class labels, we looked at 100 noun instances where these discrepancies occurred. 52 of these were identified as possessions by at least one annotator, while the remaining 48 occurred outside possession spans.

Non possessions. From these 48, 14 were nouns of type person, such as “doctor,” “dad,” “wife,” etc. which the classifier erroneously marked as possessions. Some of these were preceeded by possessive markers which the classifier most likely relied on to provide a possession label. This indicates that the signals of type person (such as those resulting from the NER module) need to be boosted through weighting or by identifying additional ways to represent them. There were also numerous instances where the target noun was part of a verbal expression such as “shocked out of our minds,” a multiword expression such as “guest house” or “big picture,” or part of a composed proper noun (such as movie names), which caused the classifier to incorrectly predict possession labels. The automatic classification also struggled with identifying possessions as they pertained to the writer of the article. We should underscore that we did encounter instances of the same noun in a single blog where one occurrence was marked as a possession in the gold standard, as it belonged to the author, while another was not, as it was talking about the same item in generic terms. One such example has a writer talking about her furniture, and then mentioning that “you can find the all weather patio furniture from Trex ”. While the classifier incorrectly distinguished between these two scenarios, it did produce opposite labels for these two instances, signaling that different encodings are indeed generated but that further refinement is needed to predict the correct class label.

Possessions. The vast majority of the possessions that were missed by the classifier were typically those that were implied or were part of a longer description, where the initial item was identified as a possession, but the subsequent items were too far removed from the context in order for the classifier to uncover signals that those were also possessions. For example, a writer talks about her yard, and follows by saying[8]:

"A little landscaping, some cute pots and plants, a few accessories and charming furniture made a world of difference!"

None of these latter possessions were identified as such. To alleviate this, we should introduce signals that also take into consideration prior possession decisions, and eventually, as was the case with the human annotators that made sure that possession annotations exhibit document level consistency, allow the classifier to make a final pass to ensure that annotation consistency is achieved. We also identified 5 instances (out of 52) that were erroneously marked as possessions by a single annotator (due to the fact that we used the bronze standard of the data); for these intances our classifier was actually correct. In the future we will aim at using the silver and gold-standard only or adding a confidence threshold to allow the classifier to model annotation strength.

6 Application

In order to gauge the impact that possession identification may have in furthering our ability of extracting latent information about users, we conduct a pilot experiment that seeks to relate possessions to author gender. We use the blog data set released by (Mukherjee & Liu, 2010), consisting of 3,226 blog posts annotated with the gender of the author. The blogs are automatically processed using the best performing possession extraction variation (22), which identifies 19,564 possession occurrences of 1,659 unique possessions. Normalized pointwise mutual information ($nPMI$) is then computed using the formula proposed by (Gerlof, 2009), where x stands for the possession word, and y for the gender class (male or female):

$$i_n(x,y) = (\ln \frac{p(x,y)}{p(x)p(y)})/(-\ln p(x,y)), \tag{1}$$

Table 6 shows the top 20 automatically identified male and female possessions as ranked by their $nPMI$ scores. While not all the entries are possessions as defined by the annotation guidelines, it is nonetheless interesting to note the preponderance of possessions from a given class within each gender group. Overall, female possessions are associated with travelling (map, passport, schedule, meetings, appointment, flight), working out (treadmill, exercise, tracks, dances, equipment), health care (dentist, appointment, mri, coverage, shelter, hospital, virus), shopping (model, apparel, store, t-shirts) and hobbies (artist, button, needle, thread, collage). Male possessions center around hobbies (pic, cigarette, saxophone, boat, tools, video, cable, tools, wire, networks, printer, ipad). Zooming in on certain topics, e.g., food, we also notice differences. For example, somehow surprisingly, male authors seem to be talking more about ingredients, exemplified by possessions such as: root, flour, mixture, salt, skillet, grill, sausage, oven, ingredients, frosting, spinach, macaroni, cake, appearing in the top 100 ranked possessions. In comparison, the culinary

[8] The goldstandard possessions are underlined.

Male possessions	nPMI	Female possessions	nPMI
pic	71.06	map	56.02
ways	63.57	dentist	56.02
owners	63.57	nails	56.02
results	55.93	potty	56.02
product	55.93	log	48.21
cigarette	55.93	rice	48.21
root	48.13	model	48.21
flour	48.13	tummy	48.21
technology	48.13	palette	40.21
saxophone	48.13	da	40.21
targets	48.13	tv	40.21
crew	48.13	missionaries	40.21
environment	48.13	horn	40.21
mixture	48.13	sides	40.21
jackets	40.19	passport	40.21
salt	40.19	priorities	40.21
boat	40.19	servers	40.21
beds	40.14	blackberry	40.21

Table 6. *Top 20 automatically identified possessions per gender as ranked by their normalized PMI scores.*

vocabulary for the female group is less specific, in the top 100 possessions, we only encounter: rice, popcorn, foods, cereal and sandwiches, yet there are more references to food related outings: servers, lunch, portions. In general, we can see that the possessions that people mention reflect not only their lifestyle, but particular stages within their life. For example, if a blogger talks about a "potty," the author is more likely a middle age female with young children. Such information can act as building blocks into defining narrower demographic slices, and establishing more comprehensive user models.

7 Conclusion

In this paper we introduced and defined the task of possession identification in text. We established an elaborate set of annotation guidelines, which enabled us to uncover 799 possessions from blog posts, with an initial moderate agreement of 44.4 percent and after reconciliation reaching 78 percent, indicating that the task is well defined. We are releasing the possession identification annotation guidelines, as well as three versions of the annotated blogs based on different confidence levels: gold-standard (where all three annotators agree), silver-standard (at least two annotators agree), and bronze-standard (containing all the annotations made), which we hope will kindle research into the area of possession identification.

We also introduced a machine learning framework to automate possession identification in text, and presented and analysed several features that can be used for this task. The classification results showed significant improvement over three different baselines. In addition, we explored the use of possession identification in an application, and used the identified possessions on a gender annotated data set to show how items that people own correlate with their gender.

Overall, as shown by the experiments we conducted, possession identification is a challenging task, yet given the results we obtained so far, we demonstrate that it can be performed automatically. In the future, we plan to craft better signals to enable improved results on this task. In addition, we will conduct more annotations to extract better patterns involving possessions. These additional annotations could be bootstrapped automatically, allowing us to annotate a very large possession dataset and to train machine learning algorithms such as neural nets that require more data to make strong predictions. We would like to also consider possessions owned by other entities than the writer, and experiment with additional machine learning features and techniques.

The annotation guidelines, the data, and the code are available for download at at `http://lit.eecs.umich.edu/research/downloads/`.

8 Acknowledgements

This material is based in part upon work supported by National Science Foundation award #1344257, by grant #48503 from the John Templeton Foundation, and by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Defense Advanced Research Projects Agency.

References

Aha, David W., Kibler, Dennis, & Albert, Marc K. 1991. Instance-based learning algorithms. *Machine Learning*, **6**(1), 37–66.

Burger, John D., & Henderson, John C. 2006 (March). An exploration of observable features related to blogger age. *Pages 15–20 of: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.*

Burger, John D., Henderson, John, Kim, George, & Zarrella, Guido. 2011 (July). Discriminating gender on Twitter. *Pages 1301–1309 of: Proceedings of the Conference on Empirical Methods in Natural Language Processing.* EMNLP 2011.

Cheng, Zhiyuan, Caverlee, James, & Lee, Kyumin. 2010 (October). You are where you tweet: A content-based approach to geo-locating Twitter users. *Pages 759–768 of: Proceedings of the 19th ACM International Conference on Information and Knowledge Management.* CIKM 2010.

Ciot, Morgane, Sonderegger, Morgan, & Ruths, Derek. 2013 (October). Gender inference of Twitter users in non-English contexts. *Pages 18–21 of: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* EMNLP 2013.

Cohen, Raviv, & Ruths, Derek. 2013 (July). Classifying political orientation on Twitter: It's not easy! *Pages 91–99 of: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.* ICWSM 2013.

Gerlof, Bouma. 2009 (Sept.). Normalized (pointwise) mutual information in collocation extraction. *Pages 3140 – 3151 of: Proceedings of the Biennial GSCL Conference.* GSCL 2009.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, **11**(1), 10–18.

Hornik, Kurt. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4**(2), 251–257.

Hu, Tianran, Bigelow, Eric, Luo, Jiebo, & Kautz, Henry. 2017. Tales of two cities: using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. *IEEE Transactions on Big Data*, **3**(1), 55–66.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation.* Chicago, IL: The University of Chicago Press.

Levin, Beth. 2006. *English object alternations: A unified account.* unpublished manuscript. Stanford, CA, USA. `http://web.stanford.edu/~bclevin/alt06.pdf`.

Li, Jiwei, Ritter, Alan, & Hovy, Eduard. 2014 (June). Weakly supervised user profile extraction from Twitter. *Pages 165–174 of: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* ACL 2014.

Liu, Wendy, & Ruths, Derek. 2013 (March). What's in a name? Using first names as features for gender inference in Twitter. *Pages 10–16 of: Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium.*

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., & McClosky, David. 2014 (June). The Stanford CoreNLP natural language processing toolkit. *Pages 55–60 of: Association for Computational Linguistics (ACL) System Demonstrations.* ACL 2014.

Mukherjee, Arjun, & Liu, Bing. 2010 (October). Improving gender classification of blog authors. *Pages 207–217 of: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* EMNLP 2010.

Nelson, Douglas L., McEvoy, Cathy L., & Schreiber, Thomas A. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 402–407.

Pennacchiotti, Marco, & Popescu, Ana-Maria. 2011 (August). Democrats, republicans and Starbucks afficinados: User classification in Twitter. *Pages 430–438 of: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD 2011.

Platt, John C. 1999. Fast training of support vector machine using sequential minimal optimization. *Pages 185–208 of:* Schölkopf, Bernhard, Burges, Christopher J. C., & Smola, Alexander J. (eds), *Advances in Kernel Methods - Support Vector Learning.* Cambridge, MA: MIT Press.

Quinlan, Ross. 1993. *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann Publishers.

Rao, Delip, Yarowsky, David, Shreevats, Abhishek, & Gupta, Manaswi. 2010 (October). Classifying latent user attributes in Twitter. *Pages 37–44 of: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents.* SMUC 2010.

Rosenberg, Milton J. 1956. Cognitive structure and attitudinal affect. *The Journal of Abnormal and Social Psychology*, **53**(3), 367–372.

Rosenberg, Milton J. 1968. Hedonism, inauthenticity, and other goals toward expansion of a consistency theory. *Pages 73–111 of:* Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T.M., Rosenberg, M.J., & Tannenbaum, P.H. (eds), *Theories of cognitive consistency: A sourcebook.* Chicago, IL: Rand McNally.

Sadilek, Adam, Kautz, Henry, & Bigham, Jeffrey P. 2012 (February). Finding your friends and following them to where you are. *Pages 723–732 of: Proceedings of the 5th ACM International Conference on Web Search and Data Mining.* WSDM 2012.

Stecher, Kristin, & Counts, Scott. 2008 (March). Spontaneous inference of personality traits and effects on memory for online profiles. *Pages 118 – 126 of: Proceedings of the second international conference on weblogs and social media.* ICWSM 2008.

Van Durme, Benjamin. 2012 (July). Streaming analysis of discourse participants. *Pages 48–58 of: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012.

Volkova, Svitlana, & Bachrach, Yoram. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, behavior and social networking*, **18**(12), 726–36.

Volkova, Svitlana, & Bachrach, Yoram. 2016 (August). Inferring perceived demographics from user emotional tone and user-environment emotional contrast. *Pages 1567–1578 of: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016.

Zamal, Faiyaz Al, Liu, Wendy, & Ruths, Derek. 2012 (June). Homophily and latent attribute inference : Inferring latent attributes of Twitter users from neighbors. *Pages 387–390 of: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ICWSM 2012.