TOWARDS AN IMPROVED LEXICON OF RELATIONAL NOUNS

ELIZABETH LEE
Department of Linguistics, University of Rochester

Abstract

The ability to detect relational nouns has shown importance in NLP tasks, particularly those that involve information extraction. For this purpose, Newell and Cheung (2018) contributed what is considered the first lexicon dedicated exclusively to relational nouns in English. In this paper, we discuss areas of improvement regarding this lexicon and propose possible steps towards an improved database for recognizing and extracting relation information from such nouns. Specific improvements that we hope to motivate are: a more encompassing list of relational nouns based on a broader definition of the term; more informative entries based on the argument structures of each noun; and, the ability to recognize and handle polysemy of nouns that yield both relational and non-relational interpretations. We hope to continue on this work in the future as well as motivate others to take these points into consideration when moving forward with tools for relation extraction.

1 Introduction

Relational nouns—nouns that denote an entity in terms of its relationship with another—are difficult for natural language processing systems to discern for a number of reasons. One is that occurrences of relational nouns often have similar syntactic behavior to that of ordinary nouns, hence they are difficult to differentiate on the basis of syntactic usage. Much evidence suggests that treating relational nouns similarly to other nouns too often yields incorrect interpretations. Another difficulty lies with the disagreement humans have over what exactly constitutes a relational noun. It has been argued that most, if not all, nouns carry the potential to denote relationships to some degree (Newell and Cheung 2018).

The ability for a system to automatically discern relational nouns would be useful for tasks involving information extraction and question answering. This goal appears to be the motivation for Newell and Cheung (2018) to create their lexicon of relational nouns. However, the decisions they made when building their lexicon raise some questions regarding the definition of a relational noun and how such a resource could actually prove useful in NLP applications.

1.1 Relational Nouns

The term *relational noun* refers to a noun that is denoted in terms of its relationship with at least one other entity. Relations expressed by relational nouns can be familial (e.g. *son, grandmother*), social (e.g. *friend, boss*), and a relative part (e.g. *edge, back*). The usage of relational nouns is demonstrated in sentences (1a) and (1b) below.

(1) a. The *friendship* between Sora and Riku has lasted over ten years.

b. Riku is Sora's *friend*.

The italicized noun in (1a) is an example of a noun which shows that there exists a relationship between two or more entities. The similar noun in (1b) denotes an entity while implying the existence of another.

As mentioned earlier, it is difficult to discern relational nouns by syntax alone as they have similar syntactic behavior to other nouns. There is one type of structure that is well-known to suggest that a noun is relational: a possessive construction of the type in which *mother* is written in the following examples. The phrase in (2a) is an example of a *prenominal possessive*. In other words, the possessor is located before the possessee in the possessive phrase and is also suffixed with *'s*. Example (2b) is conversely a *postnominal possessive*—more specifically a *postnominal genitive of* construction.

(2) a. Shanna's mother

b. The mother of Shanna

While relational nouns may often appear in these constructions, it is obvious that nouns that generally would not be considered relational often exist in identical constructions, and also that phrases of structures *x's y* and *the y of x* may not carry the same meaning.

(3) a. Michaela's cat

b. #The cat of Michaela

While *cat* alone may not necessarily be considered relational, in the context of (3a-b), there is a type of relationship being expressed. These constructions may also contribute to the thought that many nouns have the potential to have relational meaning attached. It has been suggested by Vikner and Jensen (2002) that non-relational nouns in such contexts use information called *qualia* to determine their relationship with their dependent lexical items. However, we will not be going deeper into this idea, as our focus in this paper is on nouns with intrinsic relational meanings.

Notice that (3b) has a phrase structure that most people would rather avoid using when the possessee is a sortal noun such as *cat*. Barker (2011) addresses how nouns without some sort of relational meaning attached are rarely used as the possessee in these constructions. With this in mind, we have perhaps our strongest syntactic clue when it comes to identifying relational nouns.

It may be useful to note that not all nouns that may be considered relational always denote a relationship, which can be demonstrated in the sentences below.

(4) a. John is a *child.*

b. John is acting like a *child.*

In (4a), *child* is being used in the sense that also denotes a relationship—in this case, it implies the existence of a mother and a father. On the other hand, the use of the word *child* in (4b) has nothing to do with relationships. Instead, it compares John's behavior with that of a young human and implies that he is immature. In this paper, however, we are not going to concern ourselves with identifying relational nouns in non-relational contexts like figurative speech.

It appears that semanticists have come to a consensus that relational nouns should be treated as argument-taking nouns, similarly to how verbs take arguments (de Bruin and Scha 1988, Barker 1995, Meyers et al. 2004, Asudeh 2005, Staroverov 2007, Newell and Cheung 2018), and thus have argument structures. Hence, we can demonstrate relations in a logical form as in the following examples.

(5) $[\![\text{Hannah is Karlene's friend}]\!] = friend(k, h)$

(6) a. $[\![\text{David's mother}]\!] = \lambda x.\, \lambda y.\, [mother(x, y)](d)$

b. $[\![\text{David's mother}]\!] = \lambda y.\, [mother(d, y)]$

Regarding (5), one might say that *friend* denotes the set of all pairs of entities such that one is a friend of the other ($< k, h > \in [\![\text{friend}]\!]$). Likewise in (6), *mother* denotes the set of all mother-child pairs, and furthermore, *mother(d, y)* would refer to every mother-child pair such that *y* is a mother of David.[1] After all, we do not say that someone is a friend unless they share a friendship with someone else, and we do not refer to someone as a mother unless they have had a child (and are female).

As lexical items that take arguments, they may carry additional information that so-called sortal nouns do not. For various NLP systems, these arguments can provide valuable information to extract. Therefore, the identification of relational nouns—among other argument-taking nouns—would be considerably helpful towards this cause.

The relationship between a relational noun and its arguments is inherent in its meaning. For instance, upon seeing the word 'child', the existence of a mother and father comes to mind, and by extension, perhaps grandmothers and grandfathers as well, and so on. This fact was perhaps one of the leading factors in deciding that a simple lexicon would be a useful tool in identifying relational nouns for NLP applications. In other words, by merely identifying a word as 'relational' judging by its lemma, we are able to obtain valuable semantic information.

[1] If we are only referring to biological mothers, from what we know about how biology works in our universe, there would only be one pair. However, if we were to include adoptive mothers in this set, there may be at least two.

As far as we are aware, however, there is more to the meaning behind a relational noun that must be known in order to extract relevant and valuable information. This has largely to do with the exact argument structure of a relational noun.

1.2 How to Spot a Relational Noun

From here, our task becomes especially tricky as we come up on the task of explaining what exactly defines a relational noun. As brought up before, humans find it difficult to agree on where to draw the line between 'relational' and 'non-relational'.

Let us first examine the identification of relational nouns based on semantic category. Three categories of nouns that seem to be universally accepted as relational are kinship, social, and part-of terms. Other groups mentioned across the literature include body parts (e.g. *heart, eye, leg*), properties (e.g. *speed, color, fear*), and nominalizations of verbs (e.g. *disagreement, rejection*).

1.2.1 Types of Relational Nouns

The different categories of relational nouns that we will go over in this section can be viewed in Table 1. Most of these categories have been drawn from Newell and Cheung's (2018) classification of relational nouns, although their categorization was intended to ease the annotation process.

Table 1. Prominent categories of relational nouns that we have identified within the literature; category names listed in bold identified by Newell and Cheung (2018).

Category	Examples
Kinship	mother, brother, sister-in-law
Social	friend, president, teacher
Body parts	eye, ear, nose
Relative parts	edge, middle, back
Properties/function	weight, color, fear
Nominalizations	collaboration, disagreement, refusal
Memberships	brotherhood, friendship, fellowship

The go-to examples for relational nouns often lie in the category of **kinship terms**. These include words such as *mother, father, sister,* and *brother*. Another common category is (non-kin) **social terms**, where non-kin related nouns like *friend, teacher, boss,* and *mayor* belong. Also among the generally accepted relational nouns are **relative parts** (also referred to as *part-of* or *part-whole* terms in the literature). These include the words *corner, middle,* and *end*. From here, we explore the nouns with questionable status as being relational.

The category of **body parts** is a subject of debate in the realm of relational nouns. Some insist that these words should be considered relational since they seem to fall into the broader category of part of terms (Lichtenberk, Vaid, and Chen 2011). The fact that many body part words are

polysemous and hence can have both relational and non-relational readings has been used as an argument both for (Laczkó, Butt, and King 2009) and against (Newell and Cheung 2018) their status as relational nouns. To illustrate their point, Newell and Cheung provide an example involving the word *ear*—if someone were to say, "I am drawing an ear," one would not associate this with the possessor of said ear. This ear may easily be interpreted as standing on its own.

Another category that has been identified as possibly relational nouns could be referred to as **properties**.[2] Nouns like *ability, color,* and *fear* are interpreted as relational because the attributes to which these nouns refer all belong to a possessor, thereby implying a relationship. de Bruin and Scha (1988) put focus on a subcategory of properties called *function nouns*, which are defined as referring to exactly one entity for every argument. Examples of such nouns include 'length', 'speed', 'distance', and 'rating'.

An additional category of nouns that some would argue are relational is a set of particular **nominalized verbs and adjectives**. These include words like *disagreement, forgiveness,* and *discussion*. Since it has been shown that their verb forms—*disagree, forgive,* and *discuss* respectively—take arguments that reveal relationships, this would surely parallel with relationships expressed by their nominalizations (consider the verb *befriend* and its nominal counterpart *friend*, both which clearly express relations). Looking closer at the word 'disagreement', for example, for such a thing to exist, it must be realized by at least two entities, and those two entities would be linked together through this abstract concept. Newell and Cheung (2018) leave such nominals out of their lexicon because the noun merely denotes a concept, not one of the entities in the relation.

Now, here we bring up the observation that words like *brotherhood* and *friendship* are excluded from the lexicon for the same reason explained above, yet these particular examples would appear to belong to the kinship and non-kin social categories, respectively. Perhaps this suggests that either 1) *friendship* belongs to a different semantic category, 2) that nouns can belong to more than one category, or 3) that we are actually looking at this in the wrong way—we may be focusing too much here on categorizing nouns based on their surface meanings rather than the underlying information behind them. While it is possible that nouns in particular semantic categories happen to share similar structural properties and occur in similar dependency relations, it would seem wise first to investigate what these properties are.

1.2.2 Structure of Relational Nouns

Most linguists who have studied them believe that part of the definition of a relational noun is that it takes exactly two arguments: itself and the entity to which it is related. However, different interpretations exist. Asudeh (2005), for instance, says that arguments of relational nouns are optional while, on a related note, the number of arguments that relational nouns take can vary (Meyers 2007a).

Some efforts strongly demonstrating that relational nouns can be defined in terms of their argument structures are the ones put towards the development of NomBank (Meyers et al. 2004). While the central motivation was to add more detailed annotations to the Penn Treebank, they were able to present argument structures of over 8,000 nouns, including hundreds of relational nouns, when they compiled the dictionary NOMLEX-PLUS (Meyers 2007b). The differences in

[2] These have also been referred to as *attributes*.

arguments for each noun entry suggests that this could potentially be a way either to discern relational from non-relational nouns or to split relational nouns into different categories.

While their contribution instead relates to cognitive processing, Gentner and Kurtz's (2005) discussion of *relational categories* suggests that relational nouns can be categorized based on what they take as arguments. In one example of theirs, they consider *robbery* to be a "relational schema category," which takes "relational role categories" as arguments (Gentner and Kurtz 2005). Such arguments, which for *robbery* were *thief, goods,* and *victim*, express relations between each other.

We can at least draw from these perspectives that argument-taking is a property of relational nouns, and that the ways in which they take arguments could be used to identify and categorize them.

1.3 Outline

In 2018, Newell and Cheung created what is perhaps the first lexicon devoted entirely to relational nouns in English. While their contribution is a modest effort towards overcoming the challenges of detecting relational nouns, it possesses a number of flaws. These issues are largely what motivated our project.

The goal of this project is to attempt to improve on Newell and Chueng's (2018) lexicon of relational nouns. We do so by first examining their lexicon and discussing the problems associated with it. Then, we propose solutions to these issues, hopefully to be explored and implemented in future work. Finally, we use the resources at our disposal to run several pilot studies and demonstrate one direction in which we might move forward.

The outline for the remainder of this paper is as follows. We will first spend the next two chapters reviewing Newell and Cheung's lexicon of relational nouns, critiquing the approach taken to build it, and proposing solutions to the issues that are evident. The chapter that follows will be a review of relevant works and how their contributions have inspired and differed from ours. From there, we will go into the studies and methods we used to build our lexicon and what we managed to achieve in the end. We will conclude with an improved understanding of how to identify and interpret relational nouns in the context of natural language processing.

From here until Chapter 5, the phrase *the lexicon* will refer to Newell and Cheung's (2018) lexicon of relational nouns.

2 The First Lexicon

In this chapter, we will examine in detail the creation of the first lexicon of relational nouns (Newell and Cheung 2018) and discuss the shortcomings that motivated our investigation.

2.1 Summary

Understanding that a tool for recognizing relational nouns could be useful in NLP applications like information extraction and question answering, Newell and Cheung (2018) developed a lexicon of relational nouns. Each of the 6,224 nouns in the lexicon was extracted from Gigaword and has one of three class labels: "usually relational," "occasionally relational," and "almost never relational." Of these nouns, 1,446 were labelled as either "usually" or "occasionally" relational.

Newell and Cheung (2018) explain that a noun is relational if it is defined by virtue of how it relates to another entity. When considering which nouns to define as relational for their lexicon, they decomposed the definition of a relational noun by establishing two criteria. The first is that all relational nouns must include one of the entities in the word. For example, nouns like *brother* and *CEO* meet this requirement while others like *disagreement* do not. The second requirement is that the noun must illustrate a relation in its meaning and cannot stand alone. This restriction allows for familial terms such as *mother* but not body parts like *heart*, as one does not necessarily think of this as related to another entity. These criteria were designed to reflect the actual definition of a relational noun and exclude nouns from the lexicon that should not be considered relational under this definition.

Using their refined criteria, they created a classifier by first having a group of annotators—three experts and 13 non-experts—label nouns as belonging to one of the three categories. Whenever there was a tie from a disagreement, the label would default to "occasionally relational." These annotations were used to bootstrap a classifier. This classifier marked the remaining unlabeled nouns as belonging to one of the three classes.

Their baseline classifier detected relational nouns solely on the basis of being found in possessive constructions. They also developed a feature-rich model that included dependency tree, morphological, and semantic features, as well as several hand-crafted features that included information regarding surrounding lemmas and POS tags. Through feature ablation, they determined that all features except for the morphological ones contributed to the best performance. Of the learners they experimented with, the support vector machine outperformed logistic regression, naive Bayes, and random forest models in terms of average precision and F1 score.

Table 2. An excerpt from Newell and Cheung's lexicon (2018).

Noun	Label
Frenchman	+
Frenchwoman	-
frequency	-
freshman	0
friar	-
friend	+
friendship	-
fries	-
front	+

2.2 Issues

To begin, let us point out the inaccuracies and vagueness that appear in the lexicon by examining the excerpt provided in Table 2.

It seems noticeably strange that *Frenchman* is labeled as relational while *Frenchwoman*, boasting a similar sense, is not. We do see that the rest of these words are labeled rather accurately

in accordance with their criteria specifically that *friend* and *front* are "usually relational" while *frequency* and *friendship*, while debatably relational, are labeled as "almost never relational."

Then we come to the word *freshman*, which was perhaps given the "occasionally relational" label either because there was strong annotator disagreement or because its status as relational may depend on its context. Compare the use of this word in the sentences *"When I was a freshman in high school, I got suspended for stealing a bathroom pass"* and *"University of Rochester freshman John Smith made campus-wide fame today by uploading a selfie of himself sitting beside the so-called 'Quad Fox'."* We will discuss the use of this label more in depth in Chapter 3.

Moving beyond that, as we discussed briefly in the introduction, what is the extent to which a lexicon of relational nouns would be useful in NLP applications? We are already aware that in general, a lexicon would be helpful in identifying relational nouns due to the fact that they would be difficult to detect based only on syntactic clues. Given what we know about the nature of relational nouns, we will assume that a lexicon applied to assist in NLP tasks involving relation extraction should be able to 1) identify a noun used in a given sense as relational, and 2) provide the proper argument structure for identifying the relations regarding said noun. While the lexicon partially fulfills point 1, the class labels do not provide enough information on the relational status of the individual senses of each noun, and an NLP system would have no way of properly identifying arguments of the relational nouns using the lexicon alone. Newell and Cheung do not provide any exact explanations of how they intended their lexicon to be applied for NLP tasks.

We should also review their criteria for limiting the definition of a relational noun. Why did Newell and Cheung do this, and does their new definition actually make sense in accordance with the existing literature? To reiterate, they decompose the definition of a relational noun such that it must denote one of the entities in the relationship and must also provide lexical evidence that the relationship exists. This excludes a number of nominal categories, not limited to body parts and properties. It is confusing why, for instance, they include words like *name, birthday,* and *description* while they leave out other nouns that denote properties, such as *distance*, *shape*, and *posture*.

Their motivation for this was to only include nouns that, standing on their own, heavily invoke the notion of a relation. However, if our goal is to build a lexicon to improve the quality of information extraction, we believe that it would be best to be able to identify any words that commonly denote relations.

Perhaps the lexicon is designed not as a final product, but as a work to be expanded upon or to be paired with other tools. They do state that they intend for this to be a starting point. This contribution consists of a list of 6,224 nouns, each with one of three labels reflecting how 'relational' it is, and they additionally offer their classifier to anyone who wishes to expand this list with the same labeling mechanism. It could indeed be the case that the lexicon is either meant to be a simple starting point or a companion tool in the larger task of extracting relations. Regardless, our contributions revolve around how to move forward with and improve upon this lexicon.

3 Resolving the Shortcomings

This chapter is devoted to considering the shortcomings of Newell and Cheung's relational nouns lexicon (2018) in Chapter 2 and discussing the ways we can reconsider each aspect and resolve the issues to create a better tool for identifying relational nouns.

3.1 Must We Redefine 'Relational'?

Let us begin by clarifying that we consider the task of confidently identifying a relational noun to be an unsolved issue. While the formal definition is well-known as a noun that is defined in virtue of its relationship with another entity (Barker 2011) and certain classes of nouns seem to be universally considered relational, some researchers analyze other types of nouns—for instance, body parts—and explain whether or not these should truly be considered 'relational' based on their own interpretation of the definition as well as the perceived nature and use of these nouns.

As mentioned in the previous chapter, Newell and Cheung (2018) 'operationalizes' the definition of a relational noun for the purpose of building their lexicon. Although it appears from their criteria that they are excluding entire categories of nouns that may be considered relational by some semanticists, they draw these lines based on evidence from existing literature if not from their own judgments.

One of our contributions is to provide our own boundary between 'relational' and 'non-relational' for the purpose of more fully capturing the semantic category of relational nouns and eventually creating an algorithm that will be able to accurately identify such nouns. Using this algorithm, we could take the next step in developing a more thorough lexicon of relational nouns.

3.1.1 Is Newell and Cheung (2018) Right?

Perhaps the best way to evaluate Newell and Cheung's (2018) classification of relational nouns is to review the literature and discuss how their decomposition compares with what seems to be generally accepted as a relational noun.

It appears that most if not all authors would agree that kinship and social terms should be considered relational. Most also seem to accept general part-of terms to be relational, like *edge* and *middle*. Looking back at our analysis of categories in the introduction, we are aware that some authors either show that they consider nouns in other categories to be relational with little to no explanation, argue for a particular group of nouns to be considered relational, or simply do not mention nouns in these categories as examples.

When Laczkó et al. (2009) discuss the inclusion of body part terms into the relational nouns class, they explain that words like *hand* and *eye* express part-whole relations between the part and the body, and that when these words are used in utterances such as "*John raised his hand*," the most probable interpretation is that the hand in question is a part of John. In another case, Gentner and Kurtz (2005) seem to argue for nouns like *bridge* to be considered relational, as one cannot have a bridge without a starting point and an ending point.

When taking the literature into consideration, Newell and Cheung's (2018) decomposition seems to be quite valid. They succeed in identifying nouns that should and should not, in theory, be considered relational through the use of simple criteria. However, we also believe that it is important to take into consideration nouns that still express relations regardless of how they meet their criteria, as the relations they have with their arguments can provide useful information.

3.1.2 Argument-taking Nouns and Relational Nouns

The underlying problem that motivates the creation of tools to identify such lexical items is that relational nouns, which may contain valuable information regarding relations, are difficult to automatically detect while relying only on syntactic clues, as they behave the same way syntactically as sortal nouns. What we want to know from here is why we are focusing specifically on relational nouns rather than on other argument-taking nouns as well.

The categorization of nouns in NOMLEX-PLUS (Meyers 2007a, 2007b) suggests that the fundamental difference between relational and sortal nouns is the type of arguments taken. Specifically, the ARG0 of a relational noun is itself. Some nouns that might have otherwise been considered relational are labeled differently because they take an additional argument. We consider this overlap along with the fact that the classification criteria for relational nouns, in their own way, align with the traditional definition—similar to how Newell and Cheung (2018) established theirs.

3.1.3 How to Determine What is Relational

Taking the details in this section into consideration, we describe here what we believe constitutes a noun as relational.

We consider **all argument-taking nouns** to be relational. The mindset behind this is that if a noun takes an argument, it can be argued that there exists some relationship between the predicate and its argument. Furthermore, regarding applications in NLP, the information attached to arguments of such nouns seems just as crucial as that of arguments of other nouns.

Argument-taking nouns as relational nouns can be separated by category based on their structure. We might have a category containing nouns that clearly meet Newell and Cheung's criteria—what we might call 'traditional' relational nouns—as well as categories for nominalizations of verbs and adjectives, part-whole terms, properties and attributes, and so on.

We do admit, however, that this definition is not without its complications. Mainly, we are aware of the fact that argument-taking can be "forced" upon any noun, and that any sortal noun acting as the possessee in a possessive construction can be given some relation to the possessor. With this fact, we come back around to the claim that most if not all nouns may possess some relational interpretation, and we end up with a new question: how exactly do we define an argument-taking noun?

This is why we try to use this basic guideline to determine if an argument taking noun should be considered to have relational meaning. We consider a nominal predicate and its potential argument, and then ask, "Does there exist a relation between these two other than ownership?" We realize that this is a rather loose guideline with potentially different interpretations, but at the time of writing, diving deeper into this puzzle is a task for which we do not have time to fully solve.

3.2 Arguments

Having established that a relational status of particular nouns ultimately remains up for debate, in this section we will examine the larger purpose of a lexicon of relational nouns and how we might better approach this in terms of how to properly identify and extract information on relations.

As stated in the introduction, it is generally agreed that relational nouns should be treated as argument-taking nouns. Knowing this much, one would think that handling relational nouns would involve knowledge on what kind of arguments they take as well as their valency, or how many arguments they take. The lexicon in Newell and Cheung (2018) presents no such information.

NOMLEX-PLUS (Meyers 2007b), a lexicon developed to help build Nom Bank (Meyers et al. 2004), contains over 8,000 entries of what they consider to be argument-taking nouns. Each lexical entry contains information about how many arguments a noun may take as well a selectional restrictions for each argument. (More details on this can be found in Chapter 4.) We believe that if a dictionary of nouns can be semi-automatically created that provides this much information pertaining to argument structure, we can build a similar tool that could be used to detect relational nouns and their arguments.

3.3 Polysemy

Before we go into this third point, we again consider how each noun is categorized in the lexicon. The meaning behind the labels *usually, occasionally,* or *almost never* relational has partly to do with how many of a noun's senses actually act as relational. For instance, *brother* is perhaps considered "usually relational" because every sense of the word denotes a member of a relation. Furthermore, *boy* may be labeled as an "occasionally" relational noun because it is perceived as relational in some senses and not others. In addition, nouns were sometimes given this median label if annotator disagreements were split equally.

To summarize, each word having one of these three class labels without regard to their senses fails to provide definite in formation on their relational status and how to treat them. Taking a look back at the word *boy*, imagine that an NLP system comes across one of the sentences in (7).

(7) a. Harold's **boy** starts Kindergarten tomorrow.

b. There was a **boy** playing in the park by himself.

The word *boy* as used in (7a) takes on a meaning similar to 'son', which is undeniably relational, whereas the same lexical item in (7b) is used in its more general sense (a male human child). Given the information provided by the lexicon, which is that 'boy' is "occasionally relational," how would this allow our system to discern between the two senses and extract the relevant information? While one strategy could treat all "occasionally relational" nouns the same way as "usually relational" ones, this would exacerbate the issue of unnecessarily searching for relational information surrounding nouns without relational senses.

If we were to create a tool that accurately identifies relational nouns for information extraction purposes, it seems clear that word sense must be taken into account instead of relying on a noun's lemma. Say that we could improve the lexicon in one way. We could obtain WordNet (Fellbaum 1998) senses of each noun and then run the annotation process on each word sense rather than on each lexical item. While this would create a lengthier process, it would perhaps produce a more coherent lexicon with labels corresponding to each sense of a noun rather than each noun in general. The downside to this workaround would be that the appropriate use of these labels would require applying some word sense disambiguation technique on the text to be analyzed to determine the sense of each word in order to identify its relational status.

In section 5, we explain why we do not attempt to handle polysemy given the resources we choose to employ. Hence we leave improvement on this particular aspect as a task for future work.

4 Relevant Works

There are few works aside from Newell and Cheung (2018) that have specifically dealt with the handling of relational nouns as they relate to natural language processing. The ones that we have found to be most relevant will be discussed in this chapter.

4.1 Relation Extraction

Pal and Mausam (2016) describe their improvements to a system they created called RelNoun. This nominal open IE system is designed to extract tuples representing relations through nouns. Noun-mediated patterns that are likely to contain relation information include appositive constructions (e.g. Lovely Warren, the Mayor of Rochester), possessive constructions (e.g. Rochester's Mayor Lovely Warren), and compound noun phrases (e.g. Rochester Mayor Lovely Warren). Relations described by compound NPs are particularly challenging to interpret and extract, which motivates the goal and achievements described in their paper.

They were able to improve the quality of relation extraction from compound noun phrases by focusing on three resources: capitalized relational nouns, demonyms, and compound relational nouns. Extracting information from capitalized noun phrases, as it turns out, results in new types of errors, which they deal with in the following ways. The problem of organization names being misinterpreted is handled by creating a list of 160 organization words which were extracted from last words from a list of organizations on Wikipedia. Extractions where the first argument has one of these organization words are filtered out. Issues arising from demonyms are fixed by extracting information from Wikipedia and geographic websites to create and incorporate a table of location-demonym pairs. The system would check the second argument of a tuple to see if the head-noun is in the demonym table. This approach was beneficial for other patterns, not just compound NPs. Distinctions between domicile and title names are determined by a manually written list of relational nouns that denote a high position in countries, cities, etc. (e.g. *president, governor*). As for handling errors with compound noun phrases, a list of common relational noun prefixes was created and used to segment compound relational nouns.

What we learn from reading about the improvement of RelNoun is that developing a lexicon to identify words of a particular type—in our case, relational nouns—could help to improve the performance of information extraction systems.

4.2 A Dictionary of Argument-Taking Nouns

Similar to PropBank, NomBank (Meyers et al. 2004) was created to provide additional information about nouns in the Penn Treebank II corpus and to identify the argument structure of argument-taking nouns for the purpose of serving in applications such as information extraction and machine

translation.[3] The classes of nouns that can be found in this include nominalized adjectives and verbs as well as 16 others, one of which encompasses what they consider to be relational nouns. One might say that NomBank's lexicon is on track towards the ideal lexicon of relational nouns that we have been looking for.

The group used previously analyzed data on nouns and what arguments they take, along with data from ComLex, NomLex, and PropBank, to semi-automatically build a lexicon of "markable" nouns and their argument structures, grouped by noun class. Relational nouns count for one of the classes. Others identified include *attribute*, *ability*, and *partitive*.

Meyers (2007) writes in detail about relational nouns and their argument structures in a document regarding the annotation guidelines for NomBank. To summarize, a relational noun *A* is one that takes at least one argument *B* such that *A* and *B* are connected by an implied relation *R*. To put it another way, the relational noun takes the role of the predicate as well as the subject. To illustrate this, take the phrase "*Mary's brother.*" As a relational noun, *brother* can take two arguments: ARG0 and ARG1. In this sentence, ARG0 would be *brother* while *Mary* would fill the role of ARG1.

The relational nouns are split into two subcategories, and the fundamental difference between the two is the number of arguments that they can take. One is called **DefRel**, which is where kinship terms would be categorized. The idea behind nouns of this type is that they exist by definition. Mary's brother is a brother whether or not either one of them is aware of the relationship. Nouns in the **ActRel** class take up to three arguments (ARG0, ARG1, ARG2), and the basis of these is that some action must have been done for these relationships to exist. The noun 'professor' is an example of one that belongs in this category. In the phrase "*Hannah's semantics professor,*" ARG0 would be the relational noun *professor* while ARG1 is *semantics* and ARG2 is *Hannah*.

Regarding inclusion criteria, it is stated that nouns were only considered relational if they could describe a relation between people, organizations, government-bearing locations, projects, or vessels (e.g. ships, airplanes). Furthermore, as a result of each noun only belonging to one class, several nouns that might otherwise have been considered relational were not classified as such for two reasons. One reason is that they are nominalizations of verbs, which have their own category, and the other reason is that they have an argument structure that fits in the class of nominalizations and not that of relational nouns. The example given in the document was the noun *ambassador*, which is derived from the verb sense of *represent* and takes a *to* phrase as an argument in *France's ambassador to China*.

The number of nouns belonging to these relational noun classes in the NomBank lexicon is 331, which is significantly below that of Newell and Cheung's (2018) lexicon. However, the information on the nouns in NomBank focus on argument structure and word sense, providing far more information relating to each noun. It is also worth pointing out the criteria they used for classifying a relational noun, which is largely argument-based. Perhaps if a similar effort were put forth while focusing only on relational nouns, we could create a more ideal lexicon by building on the NomBank lexicon.

[3] NomBank is available at https://nlp.cs.nyu.edu/meyers/NomBank.html

5 Pilot Studies

In this chapter, we will discuss our implementation of methods that would improve upon Newell and Cheung's lexicon (2018) in some ways based on the evaluation in sections 2 and 3.

The pilot studies we describe in this chapter serve three main purposes. One is to demonstrate examples of how we can move forward from the original lexicon and show that it requires minimal effort to expand on it. Second, we are also able to detect additional relational nouns. Finally, we are able to identify specific features that are common among relational nouns, whether to confirm observations found in the literature or to discover new patterns and examine them.

While we were eager to apply our own methods to improve upon Newell and Cheung's (2018) contribution, our time became limited and unfortunately, planned portions of this project have been rendered unfinished at the time of completing this essay. Regardless, we plan to continue working on this project and finding ways to improve on the lexicon.

5.1 The Dataset of Nouns

Under the direction of Dr. Scott Grimm and with the teamwork of the Quantitative Semantics Lab at the University of Rochester, we have created an extensive dataset of English nouns. Each noun was extracted from text data provided by the Corpus of Contemporary American English (Davies 2008) and contains over 100 features of each noun as determined in part by Stanford Corenlp (Manning et al. 2014). These features include, but are not limited to, bareness, countability, and types of dependent verbs, adjectives, quantifiers, and prepositions.

We believe that the amount of data collected to form this dataset might provide enough clues to determine whether a noun is relational. Based on pilot data from prior clustering experiments, as well as the performance of Newell and Cheung's classifier, we have sufficient evidence to believe that machine learning could be used to accurately label nouns based on how often they are considered relational.

5.2 Study One

For our first study, we began by merging Newell and Cheung's lexicon with our own noun data.[4] For each noun in the original lexicon, if that noun existed in our dataset, the class label for that noun would be added as a feature value for the corresponding noun in our dataset. In total, 4,997 of the nouns in the dataset received lexicon labels.

From there, we used packages from SciKit-learn (Pedigrosa et al. 2011) to train four different models and used the resulting classifiers to label each of the nouns in our dataset that were yet without class labels. The four methods we used were Linear SVC, Logistic Regression, Random Forest Classifier, and Complement Naive Bayes. This selection was largely motivated by the learners that Newell and Cheung (2018) experimented with when building their lexicon. Reviewing the performance of each, the random forest classifier appeared to be the best performing in terms of precision and recall (see Table 3), so we present the top ten features and their

[4] Newell and Cheung's lexicon is available at http://cgi.cs.mcgill.ca/~enewel3/publications/relational-nouns-lrec-2018/

importance values from this model in Table 4. Note that when evaluating for average precision and F1 score, we treated all "occasionally relational" labels as "usually relational" ones.

Table 3. Average precision (AP) and F1 scores of Study One in percentages.

Learner	AP	F1
SVM	14.73	50.41
Logistic Regression	14.73	46.22
Random Forest	57.70	53.07
Naive Bayes	51.25	52.73

Table 4. Top ten important features by as indicated by our random forest model in Study One. Feature importance values were rounded to the nearest thousandth.

Rank	Feature	Feature Value
1	Possessed Percentage	0.043
2	Possessive Percentage	0.041
3	Appositional Modifier Percentage	0.028
4	Verb Subject Percentage	0.027
5	Verb Object Percentage	0.026
6	Prepositional Subject Percentage	0.026
7	Bare Singular Noun Percentage	0.021
8	Definite Article Percentage	0.020
9	Singular Verb Percentage	0.020
10	Concreteness Mean	0.018

As per evidence from the literature, relational nouns often exist in possessive and appositive constructions. Other features include occurring with prepositional phrases and definite articles. Although we see that verb constructions seem to be a common trend here, we wonder if this merely has something to do with the general frequency of verb phrases.

It should be kept in mind that these important features were determined based on labels from Newell and Cheung's lexicon, which were influenced by their constrained definition of relational nouns. We attempt to obtain a more pertinent set of features through our procedure in the next study.

We additionally provide a few examples of newly labeled relational nouns in Table 5 as determined by the random forest classifier. Predictably, we get some part-of (*cog*), social (*combatant*), and property (*inanimateness*) terms. We do, however, deem words like *measles* and *prude* to be incorrectly labeled.

Table 5. Ten random nouns labeled as "usually relational" by the random forest classifier in Study One.

Study 1 Relational Nouns	
Aisle	Measles
Cog	Photocopy
Combatant	Prude
Hemisphere	Seeker
Inanimateness	Trillionth

In terms of the number of nouns labeled, 137 were labeled as "occasionally relational" while 697 were labeled as "usually relational." This gives us a total of 834 new nouns with relational labels. Combined with the nouns given labels from the lexicon, we get a total of 5,831 relational nouns. We may have ended up with fewer than we started, but we are not finished here yet.

5.3 Study Two

The procedure of this study is similar to that of Study One, but different only in the labeling mechanisms. In the hopes that we could obtain new labels for nouns based on our new criteria for relational nouns, we altered the 'gold' labels in our dataset to meet this definition. To aid us in this task, we extracted the argument-taking nouns from NOMLEX-PLUS (Meyers 2007b) and, if a noun in the dataset was labeled as "almost never relational" but existed in the list of NOMLEX nouns, the label would be converted to "usually relational." Furthermore, all "occasionally relational" labels were converted to "usually relational" ones in order to create a binary classification. This way, a noun can either be considered as "generally relational" or "generally not relational." As a result of this relabeling, we began this study with 3099 "generally relational" nouns.

Once the relabeling was complete, we essentially repeated the steps in Study One—trained the models and ran the classifiers on the unlabeled nouns. Again, the random forest classifier seemed to perform the best, and performance, top features, and new relational nouns are shown in Tables 6, 7, and 8, respectively.

Table 6. Average precision (AP) and F1 scores of Study Two in percentages.

Learner	**AP**	**F1**
SVM	44.50	83.15
Logistic Regression	44.24	83.59
Random Forest	87.11	81.41
Naive Bayes	80.34	76.24

It is interesting how the average precision and F1 scores for each classifier increase when dealing with these labels. This could suggest that these reclassified relational nouns are more easily predicted based on the features we are using.

Table 7. Top ten important features correlating with relational nouns as indicated by our random forest model in Study Two. Values were rounded to the nearest thousandth.

Rank	Feature	Feature Value
1	Prepositional Subject Percentage	0.072
2	Noun Negation Percentage	0.033
3	Concreteness Mean	0.025
4	Adjective Percentage	0.025
5	Possessed Percentage	0.024
6	Numeric Percentage	0.023
7	Other Determiner Percentage	0.022
8	Adverb Percentage	0.019
9	Conjoined Percentage	0.018
10	Demonstrative Pronoun Percentage	0.016

In the top features of this round, we see that being a prepositional subject comes out on top, with a significantly higher score than the rest on the list (Table 7). While determiners and possessives make the list again, we have some interesting new features, like occurrence with noun negations, adjectives, numerals, and adverbs.

Table 8. Ten random nouns labeled as "usually relational" by the random forest classifier in Study Two.

Study 2 Relational Nouns	
Abbreviation	Lowlands
Brink	Mutiny
Cutout	Partition
Dimension	Scorekeeper
Kleenex	Vacancy

Looking at the newly labeled relational nouns in Table 8, we again see a trend of part-of, social, and property terms getting accurately labeled, as well as some amusing errors (*Kleenex*). The total number of nouns labeled as relational this time is 6,142—which is around seven times more than in the previous study's results.

5.4 Study Three

Because of time constraints, we could not produce any results from this study before this paper was finished. Regardless, we will briefly describe the procedure here as we plan to continue our efforts later on.

We will start by taking the relational nouns as labeled in Study Two and sorting them into subcategories based on the groups we mentioned in §3.1.3 using a supervised clustering algorithm. Once those clusters are formed, we would ask human annotators to evaluate the nouns placed in each category to determine the effectiveness and accuracy of the clustering. If these nouns appear to be accurately sorted, we can say that we made a step towards improving the original lexicon by incorporating some information regarding argument structure.

5.5 Discussion

5.5.1 Grammatical Features of Relational Nouns

Here we examine commonly observed behaviors of relational nouns, partly to provide some explanation for the particular features being considered significant by our models and partly to explore ways of identifying this type of noun. While some of these features are supported by existing literature, others are suggested by our analysis in this section as well as from logical implications derived from other features.

Possessive Constructions

Perhaps the most common observation regarding the behavior of relational nouns is that they are often used in possessive constructions (e.g. "*Mary's child*"). The commonly accepted theory is that relational nouns are two-place predicates, where one argument is the *possessor* and the other is the *possessee* (Barker 2011).

Appositions

Another type of construction in which relational nouns are commonly seen is appositive constructions (Pal and Mausam 2016). Appositions are two noun phrases that occur beside each other and have the same referent. Examples of appositions are provided in bold text below in (8).

(8) a. **Carol, his mother,** was out doing errands.

b. Melinda would never forgive **Jacob, the boy who ruined her life by starting a horrible rumor**.

c. She could not be separated from **her friend Jessica**.

In example (8a), *Carol* is in apposition with *his mother*, which is the appositive. Sentences (8a) and (8c) contain relational nouns in their appositional constructions, and they demonstrate how such relational nouns can exist as either the appositive or the phrase in apposition.

Determiner Phrases

Logically speaking, this should be a common feature of relational noun usage as their roles as possessees often involve the use of a possessive determiner (e.g. *my mother*).

Prepositional Phrases

After examining several instances of relational nouns in context, one may notice that their relationships are often linked by prepositions. A diverse set of prepositions can be displayed to illustrate their prevalence alongside relational nouns, as in (9).

(9) a. The principal **of** the elementary school danced.

b. She had a debate **with** her classmate.

c. There was a disagreement **between** Sam and Nate.

d. The princess was a friend **to** all animals.

Prepositional phrases are especially important to consider when determining how to extract the relations. Perhaps this is why the occurrence of relational nouns as prepositional subjects is a significant feature in both of our studies, especially in Study Two. It would be interesting to break down and analyze the frequencies of the different prepositions with relational nouns in a future investigation.

Adjective Phrases

Adjectives have the capacity to precede nouns of all classes. However, Meyers (2007) briefly points out that adjective phrases are among the several constructions where relational nouns may be expressed (as in *interim teacher*). We wonder if taking a closer look at what types of adjectives precede these nouns and which categories of relational nouns often occur with them could yield some interesting results.

Bareness

A noun is *bare* when it is not preceded by a determiner or quantifier. For instance, the noun *school* in the sentence *John went to school* is bare. Logically from our analysis, it should be the case that relational nouns are less likely to be bare since they often appear in possessive constructions, and as such are preceded by possessive determiners.

5.5.2 Limitations and Areas of Improvement

Before we conclude this chapter, we will summarize the current limitations of our studies and how we can improve on our methods as we continue this research project.

Perhaps the first flaw that should be pointed out is that the average precision scores of most of our classifiers range from not very remarkable to horrendous. We may want to investigate why they are underperforming in such a way.

A factor that could be playing into the difficulty of classifying these nouns is the relatively low number of features in our dataset. While over 60 features were used in our models, this does not compare with the vast number of features incorporated into Newell and Cheung's model. Rebuilding our dataset with more features is an option to be considered when moving forward, and perhaps may even lead to the discovery new features associated with relational nouns.

Third, we did not have enough time to let human annotators thoroughly evaluate the generated lists of relational nouns and their clusters before this paper was finished. Hence at the moment we cannot say for certain whether or not our own lexicon could be a reliable resource. Hopefully, the right evaluations will be carried out in the near future as we continue and expand on this project.

The relabeling of nouns for Study Two was automated rather than hand-annotated. One step in the revision of these procedures in the future should be to evaluate the relabeled nouns and ensure that they actually align with our new criteria for considering nouns as relational.

A final point to note is that the dataset of nouns which we used to build our lexicon does not differentiate nouns based on word sense—each entry only corresponds to a different lemma. Because of this, we are unable to move forward with our goal of labeling each sense of a noun as opposed to each lemma while relying on our particular dataset.

6 Conclusion

The ability to detect relational nouns has shown importance in NLP tasks, particularly those that involve information extraction. Following the lead of Newell and Cheung (2018), we look towards developing an improved lexicon of relational nouns by considering the shortcomings of their 2018 lexicon, using this knowledge when discussing methods for building a new one, and presenting pilot data as an example of how to take the next step in future work. Specific improvements that we hope to achieve include a more encompassing list of relational nouns based on a broader definition of the term, more informative entries based on the argument structures of each noun, and the ability to recognize and handle polysemy of nouns that yield both relational and non-relational interpretations. We hope to continue on this work in the future as well as to motivate others to take these points into consideration when moving forward with tools for relation extraction, and that whatever comes of this effort will be useful for various natural language processing tasks.

6.1 Future Work

There are a number of different directions in which future research can go. One may be to deviate from the idea of a lexicon and instead create algorithms to automatically classify nouns as relational and detect their arguments. On another note, surely someone would argue that our criteria for identifying a relational noun are too broad, and so they may want to run their own study devoted entirely to redefining a relational noun. Then, as touched on before, we would hope to eventually see an improved lexicon in terms of size, accuracy, and attention to argument structure and polysemy.

Overall, we hope that with the knowledge presented in this paper, we have provided a better understanding of the complexities behind relational nouns and how they behave for the purposes of both linguistic theory and NLP applications.

Acknowledgments

I would first like to thank Dr. Scott Grimm, my advisor for this project and a fantastic professor. Thank you for being the one to introduce me to the wonderful world of semantics. I also want to thank the Linguistics department faculty for their helpful comments during my presentation for this project.

Next, I want to thank to classmates Hannah An, Venkat Govindarajan, and Siddharth Vashishtha. I have been incredibly lucky to be part of a cohort with some of the coolest people I have met, and I could not have made it without your kindness and support (and baked goods). I would also like to thank computer science graduate students Aeshaan Wahlang and Rik Bose for their technical and writing advice.

Finally, I want to thank my family – Peter, Patricia, and Jamie Lee – for supporting me through these past two years. I love you so much.

References

Asudeh, Ash. 2005. Relational nouns, pronouns, and resumption. *Linguistics and Philosophy* 28.4: 375–446.

Barker, Christian. 1995. *Possessive Descriptions*. CSLI Publications, Stanford, CA.

Barker, Christian. 2011. Possessives and relational nouns. in *Semantics: An International Handbook of Natural Language Meaning*, 1108-1129. De Gruyter.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. url: https://corpus.byu.edu/coca/.

de Bruin, Jos and Remko Scha. 1988. The Interpretation of Relational Nouns. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, 25–32. Association for Computational Linguistics.

Fellbaum, Christiane. 1998. WordNet: an electronic lexical database. *Language, Speech, and Communication*. Cambridge: MIT Press,

Gentner, Dedre and Kenneth J Kurtz. 2005. "Relational categories". *Categorization Inside and Outside the Lab*, 151–175.

Laczkó, Tibor, Miriam Butt, and Tracy Holloway King. 2009. Relational Nouns and Argument Structure: Evidence from Hungarian. in …

Lichtenberk, Frantisek, Jyotsna Vaid, and Hsin-Chin Chen. 2011. "On the interpretation of alienable vs. inalienable possession: A psycholinguistic investigation". *Cognitive Linguistics* 22.4, 659–689.

Manning, Christopher D. et al. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. url: http://www.aclweb.org/anthology/P/P14/P14-5010.

Meyers, Adam. 2007a. *Annotation Guidelines for NomBank–Noun Argument Structure for PropBank 2007*.

Meyers, Adam. 2007b. *Those other NomBank dictionaries–Manual for dictionaries that come with NomBank*. Technical report, New York University.

Meyers, Adam et al. 2004. The NomBank Project: An Interim Report. in A. Meyers (ed.) *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31. Boston, MA: Association for Computational Linguistics.

Newell, Edward and Jackie Chi Kit Cheung. 2018. Constructing a Lexicon of Relational Nouns. in Nicoletta Calzolari (ed.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-00-9.

Pal, H. and Mausam. 2016. Demonyms and compound relational nouns in nominal open IE. *Proceedings of AKBC*, 35–39.

Pedigrosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

Staroverov, Peter. 2007. Relational nouns and reciprocal plurality. *Semantics and Linguistic Theory* 17, 300–316.

Vikner, Carl and Per Anker Jensen. 2002. A semantic analysis of the English genitive. Interaction of lexical and formal semantics. *Studia Linguistica* 56.2, 191–226.