Semantic Framework for Comparison Structures in Natural Language

Omid Bakhshandeh
University of Rochester
omidb@cs.rochester.edu

James F. Allen
University of Rochester
Institute for Human and Machine Cognition
james@cs.rochester.edu

Abstract

Comparison is one of the most important phenomena in language for expressing objective and subjective facts about various entities. Systems that can understand and reason over comparative structure can play a major role in the applications which require deeper understanding of language. In this paper we present a novel semantic framework for representing the meaning of comparative structures in natural language, which models comparisons as predicate-argument pairs interconnected with semantic roles. Our framework supports not only adjectival, but also adverbial, nominal, and verbal comparatives. With this paper, we provide a novel dataset of gold-standard comparison structures annotated according to our semantic framework.

1 Introduction

Representing the meaning of text has long been a focus in linguistics and deriving computational models of meaning has been pursued by various semantic tasks such as semantic parsing. Deep semantic parsing (as opposed to shallow semantic parsing, such as semantic role labeling) aims to map a sentence in natural language into its corresponding formal meaning representation (Zelle and Mooney, 1996; Berant and Liang, 2014). There has been a renewed interest in deeper semantic representations of natural language (Banarescu et al., 2013) in NLP community. Open-domain semantic representations enable inference and reasoning, which is required for many language understanding tasks such as reading comprehension tests and open-domain question answering. Comparison is a common way for expressing differences in sentiment and other properties towards some entity. Comparison can happen in very simple structures such as 'John is taller than Sam', or more complicated constructions such as 'The table is longer than the sofa is wide'. So far the computational semantics of comparatives and how they affect the meaning of text has not been studied effectively. That is, the difference between the existing semantic and syntactic representation of comparatives is not distinctive enough for enabling deeper understanding of a sentence. For instance, the general logical form representation of the sentence 'John is taller than Susan' using the Boxer system (Bos, 2008) is the following:

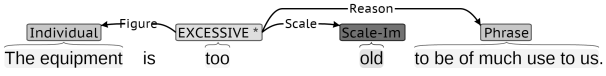
$$named(x0, john, per) \\ \& \; named(x1, susan, nam) \\ \& than(taller(x0), x1) \quad (1)$$

The above meaning representation does not fully capture the underlying semantics of the adjective 'tall' and what it means to be 'taller'. A human reader can easily infer that actually the height of John is greater than the height of Susan. Another example to consider is the sentence 'John is tall', which basically has the typical logical form $tall(john)$ –which is a very superficial representation for the meaning of the predicate 'tall'. Likewise, a human reader can infer that defining someone as 'tall' in some domain of discourse entails that this person is somehow 'taller' than some other population (say their average), however, the earlier typical logical form representation does not enable such inferences.

In this paper we introduce a novel framework for semantic representation and computational analysis of the structure of comparison in natural language. This framework enables deeper representation of semantics of comparatives, including all different types of comparison within *compara-*

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 993–1002,
Lisbon, Portugal, 17-21 September 2015.

tives, superlatives, equatives, excessives, and assetives, and the way they are related to their corresponding semantic roles. Together with this paper, we provide a dataset of gold-annotated comparative structures using our meaning representation, which enables training models on comparison constructions. We propose a new approach for automatic extraction of comparison structures from a given text. A semantic representation of the comparison expressed by the sentence 'The equipment is too old to be much of use to us.' augmented under our representation would be the following:



Throughout this paper we define a comparison to be any statement comparing two or more entities, expressing some kind of measurement on a scale, or indicating some degree of having a measurable property. The details of these variations will be discussed in Section 3.

2 Background

In this section we provide a linguistic background on comparison constructions in language, which provides the basis of our semantic framework (to be presented in Section 3).

2.1 Comparative Structures in Language

Measurement in natural language is mainly expressed in sentences having comparative morphemes such as *more, less, -er, as, too, enough, -est, etc*[1]. Comparatives can be either adjectival, adverbial, nominal, or verbal, i.e., the main component of the sentence carrying out the measurement can have either of these parts of speech.

Adjectival Comparatives: Canonical examples of comparative sentences contain adjectives, e.g., 'tall' or 'pretty'. Even within adjectival comparatives, there is a good deal of structural variety. Consider the following examples:

(1) a. Mary is *taller* than Susan.
b. Mary is 3 inches *taller* than Susan.
c. Mary is *taller* than 5 feet.

[1]These morphemes are counted as the main comparison operators. For easier representation, throughout this paper we specify the smallest constituent containing any of these morphemes as the comparison operator, which is *italicized* in the sentences.

The comparative form of the adjective 'tall' in sentence 1a is viewed as an expression denoting a *greater than ($\succ$)* relation between two individuals, 'Mary' and 'Susan', on the scale of 'tallness'. The degree-theoretic analysis of such adjectives brings up the notion of *Gradable Adjectives*: many adjectives describe qualities that can be measured according to *degrees* on *scales*, such as the scale of 'size', 'beauty', 'age', etc. These adjectives can be used with comparative morphemes, indicating less or more of a particular quality on a scale. Gradable adjectives can express specific relations between individuals on a scale, e.g., in sentence 1b Mary is taller than Susan by a measure of 3 inches.

Comparison on the scale does not always involve two individuals. For example consider sentence 1c which denotes a comparison being made between an individual and a specific point on the scale of 'tallness'. All the earlier examples are among the simplest types of comparative structures using adjectives. Consider the following example:

(2) Mary is *taller* than the bed is long.

In sentence 2 we have a case of *subcomparatives*, where we compare 'Mary' and 'bed' according to two different dimensions: height and length. Each dimension provides a degree, and the degrees are ultimately related by the *greater than ($\succ$)* relation. Scalability is known to be universal in language and a wide variety of linguistic phenomena can be explained in terms of degrees and scales (Solt, 2015).

The Semantics of Scales: A fairly common view (Kennedy, 2007) is that a scale S is a triple of the following form:

$$S = \langle D, \succ, DIM \rangle \quad (2)$$

where D is a set of degrees, $\succ$ is an ordering relation on D, and DIM is the dimension of measurement.[2]

Individuals are linked to degrees by measure functions. A measure function μ_S is the function that maps an individual x to the degree on the scale S that represents x's measure with respect to the dimension DIM. For example, the μ_{HEIGHT} measure function is a function that maps individuals to their respective heights. Under this model, we represent the comparative structure of the sentences 1a-1c as follows:

[2]To know more about the theory of scales and restrictions on dimensions and degrees refer to (Solt, 2015).

(3) a. $\mu_{HEIGHT}(Mary) \succ \mu_{HEIGHT}(Susan)$
b. $\mu_{HEIGHT}(Mary) \succeq \mu_{HEIGHT}(Susan) + 3$"
c. $\mu_{HEIGHT}(Mary) \succ 5\ feet$

A generic comparative interpretation of some degree of tallness under $HEIGHT$ scale is as follows:

$$[\![tall]\!] = \lambda d \lambda x.\mu_{HEIGHT}(x) \succeq d \quad (3)$$

where d is the degree argument which is supplied by some form of degree morpheme: a degree modifier (e.g., too, very), a measure phrase (e.g. 1.7 inches), or simply comparative or superlative morphology. Under this model, we can also represent the comparative structure of the sentence 'Mary is tall'[3], where there is no explicit degree argument. A common assumption is that the degree role is played by a phonologically null degree morpheme called pos, which denotes a context-dependent threshold or standard of comparison (Kennedy, 2007; Heim, 2007). For instance, in a specific context of adult men in north America being 'tall' could be interpreted as being over 6 feet.

Non-canonical Comparatives: Comparative structures can also be verbal, nominal, and adverbial. Consider the following verbal comparatives:

(4) a. The women ate *more* than men did.
b. The lake cooled *more* than 4 degrees.

It has been proposed (Wellwood et al., 2012) that measure functions (μ) can be applied both to individuals and to events, in the latter case measuring either the event or an entity related to the event. The comparative interpretation for the two sentences 4a and 4b is as follows:

(5) a. $\mu_{volume}(eat(women)) \succ \mu_{volume}(eat(men))$
b. $\lambda e.\mu_{coolness}(e)(lake) \succ 4\ degrees$

where cool is a function that takes an event e and an object x (here 'lake') and returns a degree representing the amount to which x changes in coolness as a result of participating in e. The underlying scale of verbal comparatives is sometimes ambiguous, e.g., in sentence 5a it is not clear whether the women ate more in volume or in quantity.

Comparative structures can also be nominal. Consider the following sentences:

(6) a. *More* juniors than seniors came to the ceremony.
b. We bought *more* milk than wine.

The meaning of sentences presented above must be stated with reference to degrees as well (Solt, 2015). Hence, the scale for the comparison sentence 6a is the numerical counting by integers and the scale for sentence 6b is something corresponding to a mass dimension, here perhaps liquid volume. Adverbial comparatives share many of their characteristics with the adjectival and verbal class, which we do not develop further for brevity. For example the sentence 'Mary ran *faster* than Sam' is an example of adverbial comparison, where the implicit 'speed' attribute of the 'running' event associated with Mary and Sam is being compared.

2.2 Categories of Comparison

There are various ways for making comparisons, each indicating different degrees of difference or similarity. Following are the major categories for degrees of comparison together with example sentences[4]:

(7) Comparative
a. Mary is *taller* than Susan.
b. Dogs are *more* intelligent than rabbits.

(8) Superlative
a. Mary is the *tallest* girl in her class.
b. Dogs are the *most* intelligent among pets.

(9) Equative
a. Mary is *as* tall as Bill.
b. Dogs are *as* intelligent as cats.

(10) Excessive
a. Mary is *too* short for basketball.
b. Dogs are *too* intelligent to be fooled.

(11) Assetive
a. Mary is tall *enough* to reach the shelf.
b. Dogs are intelligent *enough* to find their way home.

3 Semantic Framework of Comparison

As discussed earlier, having a deep meaning representation of comparison structures can help us

[3]Such cases are called *positive* usage of the adjective. The *negative* (also called antonym) usage would be 'Mary is short'.

[4]As shown in example sentences of Table 1, there can be non-adjectival comparisons in each of these categories as well.

build computational models of comparison in natural language and perform inferential tasks in various domains. Here we introduce a novel semantic framework of comparison. This framework is based on the linguistic interpretations presented in Section 2, but formalized and adapted to suit our semantic computational framework.

We model comparatives as inter-connected predicate-argument structures, where predicates are the main comparison operators (implicit and explicit comparison morphemes), and arguments are connected to the predicates via semantic roles (relations). Our framework includes not only explicit comparisons, but also implicit ones in the form of an evaluation or a measurement on a scale, which will be explained throughout this section. More detailed and complete list of the predicates, semantic roles, and arguments can be found in the supplementary material.

Predicates: Table 1 lists all the predicate operators under our framework. As the table shows, there are four main types of predicates: *comparatives, extremes, bases, and measurements*. Most of these types can be associated with operators from any of our parts of speech: Adjective (JJ), Adverb (RB), Noun (NN), and Verb (VB). The predicate operator in each of the examples is italicized. The *comparatives* type also includes the operators < and =<, which are the opposite of the operators > and >= presented in the table. It is important to note that the 'base positive' predicate is actually the implicit pos operator (as mentioned in Section 2.1; however, for easier representation we specify it by marking its corresponding adjective or adverb. The same thing happens for measurement predicates. Also, our framework captures the subtle difference between the meaning of 'Mary is $[tall]_{positive}$' and 'Mary is 5 feet $[tall]_{measurement-explicit}$'. The earlier means that Mary is tall according to some standard of tallness in a context, while the latter means that Mary's height equals the degree of 5 feet.

Semantic Roles: Each predicate is characterized by its arguments and each argument is connected to the predicates by a relation (semantic role) type. Table 2 shows the possible semantic role types for a predicate. *Figure* is the core role for a comparison structure, i.e., any comparison should have a role indicating the main entity which is being evaluated/measured/compared on a scale. The simplest form of comparative predicate, e.g., 'John is taller than Sam', involves two main roles: Figure (John) and Ground (Sam). The non-core roles are mainly associated with non-comparative comparisons

Arguments: Last but not least, each role points to an argument, which can have various types, as listed in Table 3. The most frequent argument type is individual, as in 'John is taller than Sam'. The other notable role is Phrase-value, which represents an interesting comparison phenomena. In the corresponding example in the table, the speed of John's driving is explicitly being compared with some point on the scale of speed to which 'he was allowed'. Such *ground* roles are classified as phrase-value, where a verb phrase signifies a point of comparison on scale, not an individual entity. Figure 1 shows an example of predicate-argument structure under the described semantic framework.

4 Predicting Comparison Structures

Given an input sentence, we want to predict the predicate operators, their semantic roles, and arguments. We decompose this problem into three sub-problems:

- Labeling predicate candidates using a multi-class classifier
- For each predicate, considering the set of all possible argument spans:
  - Use a classifier for predicting the role type label
  - Use a classifier for predicting the argument type label

Our overall approach, to be described in this section, is similar to the works on joint inference with global constraints for learning event relations and process structures (Do et al., 2012; Berant et al., 2014).

Predicting Predicates: The first step in comparison structure prediction is to identify and label the predicates. For this purpose we train a multi-class classifier that labels all one-word constituents in the sentence with any of predicate types in Table 1 or *None* (indicating that the constituent is not a predicate). The set of all possible predicate labels is named P.

We used various features for training the predicate classifier: we extract the lemma and POS tag of the word, POS tag of children, siblings, parent and root of the sentence in the dependency

Predicate Type	Subtype	Examples
Comparatives: Comparing against one or more entities.	>	• JJ: The car was *more* modern than I had imagined. • RB: John ran *faster* than Susan. • NN: *More* cookies than cakes were purchased. • VB: Coffee is *less* consumed than tea.
	>=	• JJ: Pizza is *as* expensive as pasta. • RB: The men ran *as* fast as the women did. • NN: That college hires *as* much professors as we do. • VB: Athletes drink as much as others.
	Superlative	• JJ: Mary is the *tallest* among her colleagues. • RB: Mike talked the *most* loudly of the group. • NN: The juniors found the *most* rock of all. • VB: The fire fighters ran the *most* among others.
Extreme: Indicating having enough or too much of a quality or quantity.	Excessive	• JJ: Mary is *too* tall to fit in the chair. • RB: Sam ran *too* fast to get caught. • NN: There are *too* many students at the party. • VB: The kid screamed *too* much.
	Assetive	• JJ: Mary is smart *enough* to accept the offer. • RB: The machine works steadily *enough*. • NN: There are *enough* professors at the party. • VB: Jack passed *enough* interviews to prove himself.
Polarity: Base form expression of +/- quality.	Positive	• JJ: Mary is *tall*. • RB:John talks *beautifully*.
	Negative	• JJ: Susan is *short*. • RB: Philip walks *slowly*.
Measurement: Indicating a measurement on a scale.	Explicit	• JJ: Mary is 5 feet *tall*. • RB: Philip is driving 60mph *fast*.
	Implicit	• JJ: Mary is *5 feet*. • RB: Philip is driving *60mph*.

Table 1: The predicate types defined under our framework.

Figure 1: A full annotation of a sample predicate-argument structure under the described semantic framework.

tree, POS tag and lemma of two adjacent words, similarity features from WordNet (Miller, 1995), word polarity features, and most importantly 'attribute concepts' for words which are adjectives (Bakhshandeh and Allen, 2015). The 'attribute concepts' are the different properties that an adjective can describe, for instance 'height' and 'thickness' are the attributes of the adjective 'gangling'. Last but not least, we include the conjunction of all these features.

Predicting Roles and Arguments: Given the predicates, one should label the predicate-argument role and predict the argument type. Here we take an approach used for semantic role labeling (Punyakanok et al., 2008): given a predicate, we collect all constituents in the sentence to build a set of plausible candidate arguments. As a result, each predicate has a set of candidate arguments which should be labeled with their argument types and be assigned with a semantic role edge. Here we jointly train two logistic regression classifiers for predicting semantic role type and argument type of a predicate-argument pair, using argument identification features from (Punyakanok et al., 2008) and using the structured averaged Perceptron algorithm (Collins, 2002). The role types can be any of the roles from table 2 or *None* (set R), and the argument types can be any

Relation Type	Description	Example
Figure	The main role being compared to something else.	- [Lara] is taller than the tree.
Ground	The main role against which the figure is compared.	- Lara is taller than the [tree].
$\text{Difference}_{\text{degree}}$	The 'plus' and 'times' roles, signifying an amount of difference on a degree.	- Sam is [twice] taller than Jim.
Domain	The explicit expression of the domain/population in which the comparison takes place	- Mary is the most intelligent among [her classmates].
Reason	The reason associated with the excessive and assetive predicates	- John is too lazy [to wake up].
$\text{Measurement}_{\text{degree}}$	The main indication of a measurement	- Henry is [5 feet] tall.
Scale	The scale on which the comparison takes place	- The [height] of the chair equals the [length] of the sofa.

Table 2: The role types under our framework.

Argument Type	Subtype	Example
Individual: An entity being compared against others.	-	- [John] is a better performer than Susie.
Reference: A referring entity, the actual antecedent of which would be resolved in discourse-level.	-	- John is 2" taller than [that].
Phrase: Introduces a degree on scale.	Value	- John was driving faster than [he was allowed].
Amount: The expression of the amount in a measurement.	Value	- Mary is [5 feet] tall. - Mary is [twice] taller than Bill.
	Very Low-Low	- Mary is [a little bit] taller than John.
	High-Very High	- Mary is [a lot] taller than Bill.
Bound: A bound/approximation being set on the amount that is expressed.	Exact	- Mary is [exactly] 5 feet tall.
	Approximate	- Sam was [about] three times faster than others.
	Lower	- John walks [at least] twice faster than you.
	Higher	- Mary is [at most] twice as smart as the others.
Scale: The scale on which the measurement is done.	Explicit	- The [height] of the bridge is too low for the van.
	Implicit	- Sam is more [available] than John is.

Table 3: The argument types under our framework.

of the ones from table 3 or *None* (set G). At the end of this stage we have two scores: $sc_{p,j,r} = log\ Pr_{p,j,r}$ where $p \in P$ is a predicate type, j is a candidate argument, $r \in R$ is a role type; and $sc_{p,j,g} = log\ Pr_{p,j,g}$ where $g \in G$.

Joint Inference: Given a sentence with its extracted predicates[5], for each predicate labeled as p, the goal is the following: find the best assignment for the indicators $y = \{y_{p,j,r} \mid p \in P, 1 < j \leq n, r \in R\}$ and $x = \{x_{p,j,g} \mid p \in P, 1 < j \leq n, g \in G\}$. Here n is the number of candidate arguments for the given predicate. We model the problem as an Integer Linear Programming (ILP). We formulate the problem as follows:

$$\arg\max_{y,x} \sum_{\substack{1<j\leq n \\ r\in R}} sc_{p,j,r}\ y_{p,j,r} + \sum_{\substack{1<j\leq n \\ g\in G}} sc_{p,j,g}\ x_{p,j,g} \quad (4a)$$

$$\mathbf{s.t.} \sum_{r\in R, r\neq None} y_{p,j,r} = 1 \quad (4b)$$

$$\sum_{g\in G} y_{p,j,g} = 1 \quad (4c)$$

$$\sum_{1<j\leq n} y_{p,j,r} = 1 \quad (4d)$$

$$y_{p,j,None} = x_{p,j,None} \quad (4e)$$

$$\sum_{1<j\leq n} y_{p,j,Figure} = 1. \quad (4f)$$

The hard constraints $4b - 4c$ each indicate a restriction on the structure of the predicate-argument relation and labels: each argument can have only

[5] A constituent is a predicate if it is labeled with any $p \in P$ and $p \neq None$.

one role and argument type ($4b - 4c$), each predicate can only have one of each role type ($4d$), a 'None' role type should be matched with a 'None' argument type ($4e$), and each predicate should have exactly one 'Figure' role ($4f$). There are also some other specific constraints such as the fact that a predicate labeled with 'comparative' cannot have a 'Domain' role type and vice verse.

5 Experimental Setup

5.1 Dataset Creation

In order to make our gold-annotated dataset we used OntoNotes (Pradhan et al., 2007) release 5.0 corpus. OntoNotes covers various genres such as conversations, news-wire, and Weblogs, which provides distinctive variations of comparison structures in natural language. Furtheremore, we think our annotations can potentially provide augmentations on OntoNotes, so using the original OntoNotes sentences can be beneficial.

One approach for pinpointing comparison sentences is to mine for some known patterns and train a classifier for distinguishing comparison and non-comparison sentences (Jindal and Liu, 2006b). However, as demonstrated earlier, the variety of comparison structures is so vast that being limited to some specific patterns or syntactic structures will not serve our purpose. In order to address this issue, we randomly selected 2000 sentences from OntoNotes which contained an adjective, an adverb, or any of the comparison morphemes. This set contained some non-comparison sentences, such as 'John admitted to the crime too'.

In order to make the final set of comparison sentences we performed the following task: we define a comparative sentence as a sentence that contains at least one predicate operator as defined in Section 3. Hence, we provided three human experts with a full predicate operator types table and asked each of them to annotate any predicate operator found in the given sentences. Then we retained any sentences with at least one predicate operator which was annotated by at least two of the three judges. We further refined the set to include equal number of predicate types. This resulted in 531 sentences.

After collecting the comparison sentences, we asked the annotators to provide gold-standard annotation of predicate-argument structure of the sentences. This involves the annotator to read the annotation guideline and basically understand the semantic framework for comparison structures that we introduced in Section 3. Initially, we ran a pilot study on a set of 50 sentences where each sentence was annotated by two of the experts. We used pilot results for iterating over the annotation schema and guideline and resolving issues regarding low agreement predicates and argument types[6], until getting to average agreement $\kappa = 0.80$. We split the dataset into 30% and 70% for testing and training respectively.

5.2 Evaluation

Here we evaluate the performance of our proposed predicate-argument structure prediction. We present the following two methods:

- **ILP Method**: Our full approach as described in Section 4. Here we used the Gurobi[7] optimization package for finding an exact solution for our ILP formalization.

- **Baseline**: A simple pattern-based method which uses lexical patterns for predicting predicate type and argument types. This method uses the generic comparative morphemes such as 'er', 'est', 'more' and 'less' for detecting any specific type of predicate. For identifying predicate arguments it relies on rules which use syntactic structure, e.g., for a 'greater' predicate identified by 'er' morpheme, the 'left' argument is always the main subject of the sentence. This method annotates anything not recognized by patterns as 'None'.

Here with compare their predictions on test set to the gold standard annotations and compute micro-averaged precision, recall and F1 score. For this analysis we remove the 'equative' predicate type, given its very low frequency in our training set. Moreover, here we do not include the positive and negative predicate types, as these take only one role argument which is 'figure', making the prediction task trivial.

Table 4 shows the results of predicate type prediction. The final reported average in this table excludes the type 'None'. The best performing category in both methods is 'superlative', which is because of its more typical structure which makes it

[6]The disagreements were mainly on fine-grained predicate types, which were resolved by collapsing some of the types together.

[7]www.gurobi.com

easier to be predicted. In general, the precision of predicate prediction is very high in ILP method, which is due to the fact that our predicates are the comparison operators indicated by the comparison morphemes. The baseline performs considerably weaker than ILP method for predicting *less* and *greater* predicates. This is because predicting these types requires a more complicated analysis where simple morphological and syntactic patterns can result in many false positives.

Table 5 depicts the results of the role type prediction. The weighted average in this table is based on frequency, excluding the type 'None'. The precision on role prediction varies across different types. Overall, the baseline performs weakly on predicting role types, which is dues to the complicated structure of roles.

The best prediction of ILP method is on scales, which has benefited from the attribute concept feature. The weaker performing types have been affected by the low-frequency occurrence in the training set. There are many cases of very long and complex sentences in our dataset. One major reason behind some of the false predictions is incorrect dependency parse for long sentences. One notable issue here is that for easier prediction and analysis, we had asked our annotators to mark only the head words for phrasal arguments. This had often caused lower agreement among annotators and hence worse predictions on the system trained on the dataset. In future, we are going to switch to span-based argument identification.

	ILP Method			Baseline		
Predicate Type	**P**	**R**	**F1**	**P**	**R**	**F1**
Assetive	100	46	63	100	26	41
Greater	90	82	86	54	68	60
Superlative	96	79	87	89	73	80
Excessive	100	43	60	100	24	38
Less	100	86	92	45	71	55
None	96	99	99	78	80	79
Average	97	67	79	77	52	68

Table 4: The evaluation results on predicate type prediction.

6 Related Work

The syntax and semantics of comparison in language have been studied in linguistics for a long time (Bresnan, 1973; Cresswell, 1976; Von Stechow, 1984). However, so far, computational modeling of the semantics of comparison components of natural language has not been developed as elaborately as needed. The main efforts on computational aspects of comparatives have been in the context of sentiment analysis. Jindal and Liu (2006b) introduced the first approach for the identification of sentences containing comparisons. Their system trains a Naive Bayes classifier for labeling sentences as comparative or non-comparative.

	ILP Method			Baseline		
Role Type	**P**	**R**	**F1**	**P**	**R**	**F1**
Plus	67	31	42	11	17	13
Ground	34	56	42	6	23	9
Scale	81	28	41	63	20	30
Figure	25	44	32	3	29	5
Reason	50	12	20	33	7	11
Domain	50	25	33	26	24	25
Times	14	50	22	30	12	17
None	97	96	96	81	78	79
Weighted Average	76	42	54	24	18	20

Table 5: The evaluation result on role type prediction.

Later works progressed into identifying the components of the comparisons: comparative predicates and arguments. For example for the sentence "Canon's optics is better than those of Sony and Nikon.", the extracted relation should be: (better, {optics}, {Canon}, {Sony, Nikon}). Jindal and Liu (2006a) detect such arguments by labeling sequential rules. Xu et al. (2011) use Conditional Random Fields (Lafferty et al., 2001) to extract relations between two entities, an attribute and a predicate phrase. These works all provide a rudimentary basis for computational analysis of comparatives, however, they lack depth and breadth as they are limited to the limited comparison structure *(Entity1, Entity2, aspect)* expressed within some sequential patterns. It is evident that the framework of comparison proposed in this paper goes beyond simple triplet annotation of comparison structures and is more representative of the linguistics literature on comparatives and measurements.

The most recent related work on comparatives (Kessler, 2014) focuses on argument identification task: given a comparative predicate, they find the arguments corresponding to it. They train a classifier for this task emphasizing on syntax information. Most of the entities in their training data are products (cameras, cars, and phones). Another recent work (Kessler and Kuhn, 2014) concentrates on the annotation of what they call multi-word predicates (such as 'more powerful', where

the comparison is not one-word such as 'calmer'). They show that annotating the modifier of comparatives (i.e., the adjectives) gives better results in classification. Both these works share the major shortcoming of the earlier works, as they are very limited to their specific patterns and fail to enable deeper representation and analysis of various complex comparative structures.

7 Conclusion

Systems that can understand and reason over comparatives are crucial for various NLP applications ranging from open-domain question answering to product review analysis. Understanding comparatives requires a semantic framework which can represent their underlying meaning. In this paper we presented a novel semantic framework for representing the meaning of various comparison constructions in natural language. We mainly modeled comparisons as predicate-argument pairs which are connected via semantic roles. Our framework supports all possible parts of speech and variety of measurements and comparisons, hence providing a unique computational representation of the underlying semantics of comparison. Furthermore, we introduced an ILP-based method for predicting the predicate-argument structure of comparison sentences.

With this paper, we provide a novel dataset of gold-standard annotations based on our semantic framework. We are planning on expanding our gold-standard annotations under this framework for having more training data. Our semantic framework on comparison constructions enables us to do logical reasoning and inference over comparatives. In the future, we are planning to design a reading comprehension task where we use this framework for answering comparison questions from a paragraph containing various interrelated comparisons.

Last but not least, the works on broad-coverage semantic parsing (Allen et al., 2008; Bos, 2008) can all benefit from our semantic framework. We will be extending the TRIPS logical form (Allen et al., 2008) according to this framework and will modify the grammar to generate the deeper representations.

Acknowledgments

We would like to thank Alexis Welwood for her invaluable comments and guidelines on this work. Moreover, we thank Ritwik Bose for his help on annotations. This work was funded by the Office of Naval Research (grant N000141110417) and the DARPA Big Mechanism program under ARO contract W911NF-14-1-0391.

References

James F. Allen, Mary Swift, and Will de Beaumont. 2008. Deep semantic analysis of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 343–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Omid Bakhshandeh and James F. Allen. 2015. From adjective glosses to attribute concepts: Learning different aspects that an adjective can describe. In *Proceedings of 11th International Conference on Computational Semantics (IWCS)*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Association for Computational Linguistics (ACL)*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Brad Huang, Christopher D. Manning, Abby V. Linden, and Brittany Harding. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Joan Bresnan. 1973. Syntax of the comparative clause construction in english. *Linguistic Inquiry*, 4(3):275–343.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Max Cresswell. 1976. The semantics of degree. *Barbara Hall Partee (ed.)*, pages 261–292.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 677–687.

Irene Heim. 2007. Little. In *In Proceedings of 16th Semantics and Linguistic Theory Conference*, Cornell University, Ithaca.

Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 244–251, New York, NY, USA. ACM.

Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1331–1336. AAAI Press.

Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, February.

Wiltrud Kessler and Jonas Kuhn. 2014. Detecting comparative sentiment expressions – a case study in annotation design decisions. In *Proceedings of KONVENS*, Hildesheim, Germany.

Wiltrud Kessler. 2014. Improving the detection of comparison arguments in product reviews. In *Proceedings of 44th Jahrestagung der Gesellschaft fur Informatik e.V. (INFORMATIK 2014)*, pages 22–26, Stuttgart, Germany, September.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Sameer Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 517–526, Washington, DC, USA. IEEE Computer Society.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist.*, 34(2):257–287, June.

Stephanie Solt. 2015. Measurement scales in natural language. *Language and Linguistics Compass*, 9(1):14–32.

Arnim Von Stechow. 1984. Comparing semantic theories of comparison. *Journal of Semantics*, 3(1):1–77.

Alexis Wellwood, V. Hacquard, , and R. Pancheva. 2012. Measuring and comparing individuals and events. *Journal of Semantics*, 29(2):207–228.

Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. 2011. Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, 50(4):743–754.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 1050–1055. AAAI Press.