A Framework for Interpreting Bridging Anaphora

Parma Nand and Wai Yeap

School of Computing and Mathematical Sciences,
Auckland University of Technology,
Auckland, New Zealand
{pnand,yweap}@aut.ac.nz
http://www.aut.ac.nz

Abstract. In this paper we present a novel framework for resolving bridging anaphora. We argue that anaphora, particularly bridging anaphora, is used as a shortcut device similar to the use of compound nouns. Hence, the two natural language usage phenomena would have to be based on the same theoretical framework. We use an existing theory on compound nouns to test its validity for anaphora usages. To do this, we used human annotators to interpret indirect anaphora from naturally occurring discourses. The annotators were asked to classify the relations between anaphor-antecedent pairs into relation types that have been previously used to describe the relations between a modifier and the head noun of a compound noun. We obtained very encouraging results with an average Fleiss's κ value of 0.66 for inter-annotation agreement. The results were evaluated against other similar natural language interpretation annotation experiments and were found to compare well.
In order to determine the prevalence of the proposed set of anaphora relations we did a detailed analysis of a subset 20 newspaper articles. The results obtained from this also indicated that a majority (98%) of the relations could be described by the relations in the framework. The results from this analysis also showed the distribution of the relation types in the genre of news paper article discourses.

Keywords: anaphora resolution, noun phrase anaphora, discourse structure, noun compounds, noun phrases.

1 Introduction

The term **anaphora** originated from an ancient Greek word "$\alpha\upsilon\alpha\phi o\rho\alpha$" which means "the act of carrying back upstream". In the context of natural language processing, the term **anaphor** is a reference which points back to a noun that has been mentioned previously in the text being processed. The referred noun is called the **antecedent**. The anaphor can be the same noun as the antecedent, a variation of the noun or a completely different noun. A common form of anaphor is one in which the anaphor is used as a co-reference pointer to the antecedent noun. This is true in the case of pronouns where the pronoun has a one-to-one relation with the antecedent. It is also true in the case of some noun phrases (NPs) where the anaphoric noun directly co-refers to the antecedent (eg. James Smith/Mr Smith). However a noun can also be used as an indirect reference to a previously mentioned noun. As an example consider the following excerpt:

1. *John bought a **house**. The **windows** are wooden.*
2. *John was bitten by a **snake** on the foot. The **poison** had gone up to the knee by the time ambulance arrived.*

In the example (1) above, the noun **windows** can only be interpreted fully in the context of the noun **house** mentioned in the previous sentence. In this case, the anaphor-noun **windows** is related to the antecedent-noun in an indirect way, different from the one-to-one co-reference type relation. The example in (2) shows another indirect relation between **poison** and **snake**, however it is not the same as the one in (1). NP anaphora resolution studies (e.g. [24, 8, 27]) treat these indirect relations as a single category and refer to them as *associative* or *bridging* anaphora. In this study we propose a relational framework that distinguishes between the different types of anaphoric relations that can exist between two nouns, one of which represents the anaphor and the second one represents the antecedent. Hence in this framework, the task of anaphora resolution involves identifying the antecedent as well as the type of relation to the antecedent. To distinguish from the previous works, we will use the term *anaphora interpretation*, instead of *anaphora resolution*, where the latter involves only identification of the antecedent.

Since we are also identifying the type of relation, it is possible for an anaphor to have multiple antecedents, related by the same or a different relation. This is a significant departure from the conventional notion of anaphora resolution where an anaphor is resolved to a single, previously mentioned entity. In the case in which the antecedent is also an anaphor, it is assumed to be already resolved, forming a sequential chain. For some NP anaphora this is inadequate. As an illustration, consider the excerpt below:

The robber jumped out of the $\textbf{\textit{window}}_1$.
The $\textbf{\textit{house}}_2$ belonged to Mr Smith.
The $\textbf{\textit{window}}_3$ is thought to have been unlocked.

If we allow a single resolution relation for an anaphoric NP, then $window_3$ would have to be resolved to either $house_2$ or $window_1$. In either case, a part of the information would not be captured. A common strategy in most studies (eg. [24, 8]) is to resolve to the most recent antecedent. In the case of the above excerpt, this would mean that we resolve $window_3$ to $house_2$ which can be assumed to be already resolved to $window_1$. There are two inadequacies in this strategy; firstly the semantic difference between the relation of $window_3$ to window$_1$ and $window_3$ to $house_2$ is approximated by a single co-reference relation, and secondly as a consequence, the direct relation between $window_3$ to $window_1$ is not captured. In the proposed framework, we will identify both $house_2$ and $window_1$

as antecedents and interpret each of them with a different relation. It can be argued that this can be overwhelming since we can form a relation even between a pair of very remote entities. However the constraint in our case is that we are only interested in **relations that give rise to anaphoric use of NPs**. The interpretation framework involves **specifying a relation** between an anaphor and the antecedent hence a consequence of this is that an NP can form relations with **more than one antecedent**. This allows us to represent and interpret anaphoric uses of a noun such as *window* to an occurrence of *house* **and** another occurrence of *window*.

Identification of the specific relations in the proposed framework also allows us a richer interpretation of anaphora which are represented by more than one word, that is, compound nouns. In this case, the framework allows us to interpret the modifier-noun with a relation, in addition to the head noun. As a simple example, the compound noun *battle fatigue*, appearing after the clause "*The battle caused fatigue*" has a co-referential relation to the noun *fatigue*, but in addition it also has some semantic relation (identified later as CAUSE) to the noun *battle*.

Hence, there are two novel aspects to this framework for interpreting anaphora. Firstly it identifies a specific relation between the anaphor and its antecedent. Secondly, it also interprets modifiers beyond using them to merely identify the antecedent for the head noun, that is, it interprets them in the same way as the head noun. A consequential effect of this is that an NP can have more then one antecedent. Thus this framework enables us to determine the relational dependence of an anaphoric NP to **all** other NPs in the discourse.

2 Related Works

NP anaphora resolution has received considerably less attention from computational linguists compared to pronominal anaphora even though the proportion of NP anaphora in natural discourses is either comparable to, or more then the proportion of pronominal anaphora. The reason for this seems to stem from the fact that the problem of pronoun resolution is much better defined compared to NP anaphora. This difference in complexity of the problem also explains why whatever published work is available on NP anaphora resolution, is predominantly focussed on NPs that are definite descriptions (eg. [24, 8, 2, 3]) with the accompanying task of identifying whether a definite NP is anaphoric or not. NP resolution in these studies involves identifying a single previously mentioned noun that the anaphoric NP refers to. Anaphora in these studies have been studied as two categories; *direct* and *associative*. The direct category includes cases in which an NP directly co-refers to another entity such as the case of *he/John*. The associative category includes cases such as *window/house*. Some of the studies such as [24] have gone a step further to specify the actual associative relation in terms of synonymy[1], hyponymy[2] and meronymy[3]. The motivation for these relations seems to have risen from organization of the lexicon, WordNet [7] which is used to bridge the meanings between the anaphor and the antecedent.

In this paper we propose a framework that presents an enhanced interpretation of the generic bridging relations. The framework is based on recognizing that anaphora is used in a way similar to another natural language phenomenon, namely compound noun generation. A compound NP of the form *noun + noun* (N + N) consists of two nouns which have some underlying semantic relations ([17, 6, 19]). According to these studies, use of compound NPs is highly *productive* rather then *lexical*. In this productive process, compound NPs are formed on the fly as a discourse is being produced, rather then recalled and used from a lexicon. In this productive process, the semantic relation between the nouns is deleted and a shortcut is formed by juxtaposing the two nouns to form a compound noun. However, for interpretation of the compound noun the semantic relation is expected to be reconstructed by the consumer ([17]). This process of compound noun generation has been described as *predicate deletion* in literature. The framework proposed in this paper is based on the premise that associative anaphora usage is a similar natural language phenomenon to compound noun generation. They both involve two nouns connected by a relation, but the relation is not explicitly expressed by the producer, rather, it is expected to be deduced by the consumer. The difference is that, in the case of anaphora, the two nouns are used separately as anaphor and antecedent, while in the case of generation of compound nouns, the two nouns are juxtaposed together as a compound noun. Research on the generation of compound nouns is at an advanced stage with various theories existing on how compound nouns are formed. According to these theories, formation of NPs is not totally unconstrained, in other words, a compound noun cannot be formed with any two random nouns. For example, *war man* can not be formed on the basis of the relation "man who hates war" or similarly *house tree* can not be formed from "tree between two houses" [29]. In both the examples there does exist a relation between the nouns, however it is of the type that can be used to form a compound NP. Linguistic studies on compound nouns (eg. [6, 29, 17, 28] have assumed that the set of generic relations are finite and characterizable, although the set is not necessarily common among all the studies. Studies such as [17] and [6] have attempted to identify these relations, and even though the exact set of relations proposed by the different studies are slightly different, a core set is very similar. An additional aspect highlighted in [6] is that compound nouns can also be formed from "temporary or fortuitous" relations, hence it presents a case for existence of unbounded number of relations although the vast majority of compound nouns fit into a relatively small set of categories [26].

The relational frameworks used in computational linguistics vary along similar lines as those proposed by linguists. Some works in the computational lin-

[1] same meaning relation

[2] same subset/superset relation

[3] part/whole relation

guistics (eg. [4, 20]) assume the existence of an unbounded number of relations while others (eg. [16, 13]) use categories similar to Levi's finite set. Yet others (eg. [22, 14]) are somewhat similar to [28]. Most of the research to date has been domain independent, done on generic corpus such as Penn Tree Bank, British National Corpus or the web.

The later works on noun compounds have followed on from either [18] or [28] with some of them coming up with a slightly different variation while others have defined a finer grained set of relations dictated by the data sets used for the study. For example, [26] reports a set of 43 relations grouped into 10 upper level categories. Most of the relations from different studies can be mapped to an equivalent relation in other studies.

For this study we chose the set of relations proposed in [18] for two reasons. Firstly, our analysis of corpus for anaphor-antecedent relations seemed to map better to Levi's set of nine relations for compound nouns and secondly more of these relations can be computationally determined from existing lexicons such as WordNet and the Web. There are already several works that extract Levi's set of relations from WordNet and the Web with various levels of success. In terms of natural language processing, a linguistic theory is only useful if it can be reasonably implemented in a computational system. The theory on anaphora proposed in this paper can be easily implemented by adopting the relation extraction techniques from compound noun generation works. We are currently in the process of developing an anaphora resolution system by integrating the various relation extraction strategies described in computational works on compound nouns.

3 Anaphora Resolution Framework

In the Introduction we stated that anaphora interpretation and noun compound generation are two indicants of the same underlying relational framework between entities. Hence, a framework describing compound noun generation has to apply to anaphora usage as well. In the proposed framework we extend the relations proposed for compound noun generation from [18] for interpretation of noun phrase as well as pronominal anaphora.

An indirect reference such as *window* referring to *house* and *diesel* referring to *truck* is based on the predicates "house **has** windows" and "a truck **uses** diesel". In the case of compound noun generation, the predicate is deleted and the two entities are juxtaposed to form the noun compounds *house window* and *diesel truck*. For interpretation of the compound noun the consumer is expected to reconstruct the relation between the modifier and the head noun ([6, 18]). We propose that the compound noun generation process is very similar to associative anaphora, except in the latter case the modifier is not necessarily bound to the head noun as part of a noun compound. That is, it may exist in another clause, however the same relation is still expected to be reconstructed for a full interpretation of the anaphor. Hence, for the example for the predicate "house has window", we could have the full NP, *house window* produced by predicate deletion. However in addition, the same predicate could also be expressed anaphorically as in the following example:

John bought a ***house*** *in Glen Eden.*
The ***windows*** *are wooden.*

In the example above the related entities from the predicate “house has windows” are separated into two different sentences, each expressing information about “house” and “windows” respectively. In order to relate the two sentences we need to bridge the “semantic gap” between “windows” and “house”. This is referred to as *text cohesion* and/or *coherence* ([12, 25]). Hence identifying the specific relation between an anaphor and its antecedent is necessary for establishing coherence which is fundamental for a full interpretation of any text.

Semantic relations between certain entities exist by default and can be assumed as part of the lexical knowledge of the consumer. For example, the HAVE relation between *car* and *tyre* is part of lexicon so the noun compound *car tyre* and the noun *tyre* used anaphorically to refer to *car* is readily understood. In addition, the HAVE relation can also be established temporarily in a discourse followed by its use for anaphora and/or compound nouns. For example, after specifying the relation “the box has tyres”, the noun *tyres* can be used to indirectly refer to *box* in the same way as the reference of *tyres* to *car*. However, the former can only be used in the context of the discourse in which the relation was expressed. This corresponds to Downing’s [6] fortuitous relations. We distinguish between these two type of relations as **persistent** or **contextual**. Persistent relations are those that form part of the lexical knowledge which are valid within the context of a particular discourse as well as all other discourses. On the other hand contextual relations are transient, and may be valid only for the duration of a single discourse, for example, “a cup on a table” or “John has a knife”. The contextual relations are expressed as either a verb or a preposition, relating two entities in the discourse being processed. In order to resolve all bridging anaphora in a discourse, we need to identify both persistent as well as the contextual relations. The persistent relations can be expressed either explicitly or assumed as part of the lexicon. On the other hand, the contextual relations have to be expressed explicitly via verbs and prepositions. The question now is which verbs and prepositions represent the anaphoric relations.

As argued earlier, the semantic relations used by bridging anaphora are the same as those used for compound noun generation, hence for this study we adopted the set proposed in [18]. The set of relations consist of CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM and ABOUT. In order to define a complete framework for anaphora interpretation, we needed to do two modifications to the nine relations from [18]. Both of these modifications were done in order to be able to better interpret and represent plural anaphoric nouns. This was done by introducing a new relation named ACTION, and by splitting the existing BE relation into BE-INST and BE-OCCR. These are explained next.

When two or more entities in a discourse are participating in the same or similar event, they can be referred to as a unit by a collective NP in the context of the discourse. The entities in the same or similar action can be expressed by the conjunction *and* or described by two different clauses. For instance in the sentence “The coastguard and Lion Foundation Rescue helicopter were called out.”, the entities *coastguard* and *Lion Foundation Rescue helicopter* are related to each other by the virtue of participating in the same action. Similarly, the clauses “the truck rolled down the hill” and “the ball rolled down the hill” would enforce the same relation between *truck* and *ball* since they are both engaged in the same action (roll). This relation between truck and ball is only valid for the context of the discourse, hence this relation is contextual. We describe this contextual relation as the ACTION relation which relates entities participating in events which are identified to be same or similar. The ACTION is used to describe an NP such as *runners* used to refer to *fox* and *Peter* from the context clause “The fox and Peter were running”.

The second modification involved defining a finer grained BE relation in order to interpret existence of plurals in a different form. We split Levi’s BE relation into BE-OCCR and BE-INST to distinguish between direct co-reference or identity relation and an instance relation. In a BE-OCCR relation an NP directly forms a one-to-one co-reference to another NP, eg. *John/he* and *John/the driver.* The BE-INST relation represents cases where an anaphor refers to a plural antecedent, in a partial capacity, for example, *both trucks/northbound truck.* In this case the NP *northbound truck* is an instance of *both trucks* which is distinct from a co-reference relation. It can be argued that all subset/ superset relations such as *John/driver*(John is an instance of driver) and *car/vehicle* (car is and instance of vehicle) is an instance relation. However we consider these as BE-OCCR relation since they **function** to identify the entity. Hence in the framework, the BE-INST relation only relates a plural NP and an NP representing a subset of the plural NP.

With this discussion we can now define and exemplify the eleven relation types used in the anaphora interpretation framework. They are:

CAUSE - Includes all causal relations. For example, *battle/fatigue*, *earthquake/tsunami*

HAVE - Includes notions of possession. This includes diverse examples such as *snake/poison*, *house/window* and *cake/apple.*

MAKE - Includes examples such as *concrete house*, *tar/road* and *lead/pencil.*

USE - Some examples are *drill/electricity* and *steam/ship.*

BE-INST - Includes plural cases such as *both trucks/southbound truck*, *John/teachers.*

BE-OCCR - Describes the same instance participating in multiple events. For example *John Smith/Mr Smith/he* and *John Smith/the driver.*

IN - This relation captures grouping of things that share physical or temporal properties. For example *lamp/table* and *Auckland/New Zealand.*

FOR - This includes purpose of one entity for another. For example *pen/writing* and *soccer ball/play.*

FROM - This includes cases where one entity is derived from another. For example *olive/oil* and *wheat/flour*.

ABOUT - Describes cases where one entity is a topic of the other. For example *travel/story* and *loan/terms*.

ACTION - This is only a contextual relation meant to capture entities engaged in same or similar action either with the same object/s or a null object.

The next section describes the annotation experiment done in order to validate that anaphora usage is based on the above relation types.

4 Annotation Experiment

4.1 Annotators

For the purpose of human validation of all relations in the framework we used second and third year students enrolled in computer related degrees. The annotation experiments were done over a period of 4 weeks at the beginning of their usual classes. Four different streams were used each consisting of approximately 30 students. The students in each stream were given a basic training on the requirements of the annotation and they were given a single annotation task at the beginning of their class over a 4 week period. The whole annotation experiment was broken down into session based tasks involving 25 anaphoric NPs per task. This was done to ensure that each task was completed in about 10 minutes with minimal impact on the students class time. In addition, the annotators were not identified in any of the tasks. We only ensured that an annotator did not annotate the same task twice.

4.2 Annotation Data

Our base input data used for content analysis for all aspects of NP usage consisted of 120 articles (of mixed genre) from *The New Zealand Herald*, *The Dominion Post* and *The Press* which are three major online newspapers from three different cities in New Zealand. The choice of the articles were not completely random. This corpora was developed to serve as the input data for the anaphora resolution system which is the parent project of this study. Hence, the corpora was developed from the articles which were not too short (had more then 20 sentences), exhibited use of a variety of anaphoric uses (including pronominal anaphora) and had been written by different writers.

An inherent challenge in most NLP tasks is what is referred to as *data sparseness.* The term is used to describe a characteristic when a single chosen corpus cannot be used for consistent empirical validation of **all** aspects of a theory. This is because the prevalence of the different characteristics of an NLP theory can be unevenly distributed in a fixed corpus. Hence, we searched an extended corpus in order to make a lower threshold of 15 relations from each category. For this we used The Corpus of Contemporary American English [5]. This freely available corpora consisting of some 410 million words from a variety of genre and has an online web interface which can be used to do fairly complex searches for words and phrases hence forms an excellent resource for manual content analysis for NLP tasks.

We excluded validating the BE-OCCR relation since this is a non-ambiguous co-reference relation.

For the annotation experiment we used 3 streams of approximately 30 students giving us a total of 90 different annotators. Each annotator took 4 different tasks, one per week for a period of 4 weeks. Each task consisted of 25 antecedent-anaphor pairs and was annotated by 2 streams, ie. approx. 60 annotators. We randomly discarded some annotation task sheets in order to have a consistent number of annotations for each pair resulting in 25 annotators for each task. Each relation type from (CAUSE, HAVE, MAKE, USE, BE-INST, IN, FOR, FROM, ABOUT, ACTION) as classified by the author was represented by 15 anaphor-antecedent pairs. The pairs from each of the 10 relation types were randomly selected to make up 6 task sheets, each consisting of 25 pairs. The total number of classifications for all relations amounted to 3750 with 375 classifications for each relation type consisting of 15 different anaphor-antecedent pairs.

Each of the streams were given a basic training on semantic interpretation of the relation types using the examples in section 3. These examples were also given as a separate sheet with each annotation task. Each task sheet consisted of anaphor-antecedent pairs and a tick box for each of the relations. The annotators were asked to choose the relation which best describes the anaphor-antecedent pair. Two additional options, OTHER and NONE were also given. The OTHER was to be used if the annotator thought that a relation does exist but is not present in the given list and option NONE to be used if the annotator thought that the pair were not related at all.

	CAUSE	HAVE	MAKE	USE	BE–INST	IN	FOR	FROM	ABOUT	ACTION	OTHER	NONE
CAUSE	**208**		45	87				4	14		15	2
HAVE		**196**	113	7		13		46				
MAKE		45	**206**	120				4				
USE		45	26	**242**			59	3				
BE-INST					**347**			19			9	
IN		64	37			**241**		33				
FOR		18	5	132			**216**	4				
FROM		9	17			35	87	**227**				
ABOUT	48	11		56				7	**253**			
ACTION									5	**351**	19	

Table 1. Confusion Matrix for the non-normalized NPs. The columns give the annotations by annotators against the author's annotations on the rows. Each relation category had a total of 375 annotations done by 50 different annotators. The bolded entries indicate number of annotations agreeing with author's annotations

Table 1 shows the confusion matrix of the relation types as identified by the annotators against the author's classification. Table 2 shows the corresponding

	CAUSE	HAVE	MAKE	USE	BE–INST	IN	FOR	FROM	ABOUT	ACTION	OTHER	NONE
CAUSE	**0.55**		0.12	0.23				0.01	0.04		0.04	0.01
HAVE		**0.52**	0.30	0.02		0.03		0.12				
MAKE		0.12	**0.55**	0.32				0.01				
USE		0.12	0.07	**0.65**			0.16	0.01				
BE-INST					**0.93**			0.05			0.02	
IN		0.17	0.10			**0.64**		0.09				
FOR		0.05	0.01	0.35			**0.58**	0.01				
FROM		0.02	0.05			0.09	0.23	**0.61**				
ABOUT	0.13	0.03		0.15				0.02	**0.67**			
ACTION								0.00	0.01	**0.94**	0.05	

Table 2. Confusion Index Matrix between the relation types corresponding to table 1. The bolded entries indicate the index of annotations which were the same as that of the author's.

confusion indices between the relation types. The confusion indices indicate the likelihood of a relation type to be interpreted as another type.

4.3 Annotation Results

The first observation of the annotation results from table 1 is that only 2 annotations out of a total of 3750 were classified as NONE indicating that the annotators by and large thought that the pairs given had **some** relation. In addition a total of 43 (approx. 1.1%) annotations were classified as OTHER, which represents the relations which were described by a relation not in the list of 10 that were given. The main categories that were interpreted as having some other relation were CAUSE and ACTION, however these were still a very small percentage with indices of 0.04 and 0.05 respectively. The bolded entries in table 2 give the percentage agreement of the relation types agreeing with that of the author. The relation types BE-INST and ACTION have the highest conformance indicating they are the least ambiguous. The other types vary from a low figure of 0.52 for HAVE to 0.67 for ABOUT, with an overall agreement value of 0.66. The relation types that were easily confused and hence can be interpreted as semantically close, were HAVE, MAKE and USE. Conflating these 3 categories gives us an agreement index of 0.89. Another crucial observation is for the FROM relation. Although not by large amounts, this relation type seems to be confused with all other categories. This prompted us to closely examine the task sheets to see if there were consistent misclassifications by the author, however no such patterns were found. Some of the classifications seemed to use the FROM type as a “fall back” category.

In order to compare the inter-annotator agreement with other similar studies we also computed the Fleiss' κ measure. The κ index for the overall annotation tasks was computed to be 0.64 and the value with HAVE, MAKE and USE conflated was 0.86. The overall κ value 0.64 compares well with the inter-annotator figures from other annotation experiments dealing with identification of relations. For comparison some of the results are summarized in table 3.

Study	Agreement Index	No. of Relations
[26]	0.57 - 0.67 κ	43
[9]	0.61 κ	22
[23]	0.68 κ	6
[14]	52.31 %	20
[10]	0.58 κ	21

Table 3. Inter-annotator agreement comparison between studies dealing with relations between composite nouns of noun compounds.

5 Prevalence of Anaphoric Relations in News Articles

In order to gauge the prevalence of the anaphoric relations in naturally occurring discourses, we used 30 of the 120 newspaper articles used for the annotation experiment and analyzed them in detail to determine the existence and the distribution of the proposed relations. The set of 30 articles consisted of 352 sentences and 2323 nouns. The 30 randomly chosen articles were analyzed by the author for the existence of the 10 relations from table 1. In addition there were two additional relation types. The first one was the OTHER relation for relations that could not be categorized into any of the 10 types from table 1. The second one was the BE-OCCR relation which was considered trivial, hence was not tested in the annotation experiment. This represents the identity or the co-reference relation.

Out of a total of 2323 nouns, 1324, or 57% were found to be used anaphorically. This shows that more than half of the nouns used were anaphoric, hence highlights the importance of being able to resolve them for discourse interpretation. Note that in our framework, an anaphor can have more than one antecedent where the antecedents are related by different relations. The 1324 nouns used anaphorically had a total of 1588 relations between them. This gives us an average of 1.2 relations per anaphor. The detailed distribution of the relation types are shown in figure 1. The figure firstly shows that the majority (524) of the relations are of type BE-OCCR which are identity relations represented by both pronouns as well as noun phrases. The reset of the relations were fairly evenly distributed ranging from 64 to 175. Only a small number, 32 or 2% of the relations were found to be outside the range of the relations in the framework. Aside from the BE-OCCR and OTHER relation types, there were 1032 bridging relation from the list BE-INST, CAUSE, HAVE, MAKE, USE, IN, FOR, FROM, ABOUT and ACTION. This means a substantial proportion (65%) of relations were bridging, highlighting their prevalence in news paper articles. We are in the process of implementing the resolution of these types of bridging as well as the traditional co-reference anaphora at a discourse level. The resulting network of relations between nouns in the discourse will provide us with an infrastructure which can be utilized in a computational system for a richer interpretation of discourses.

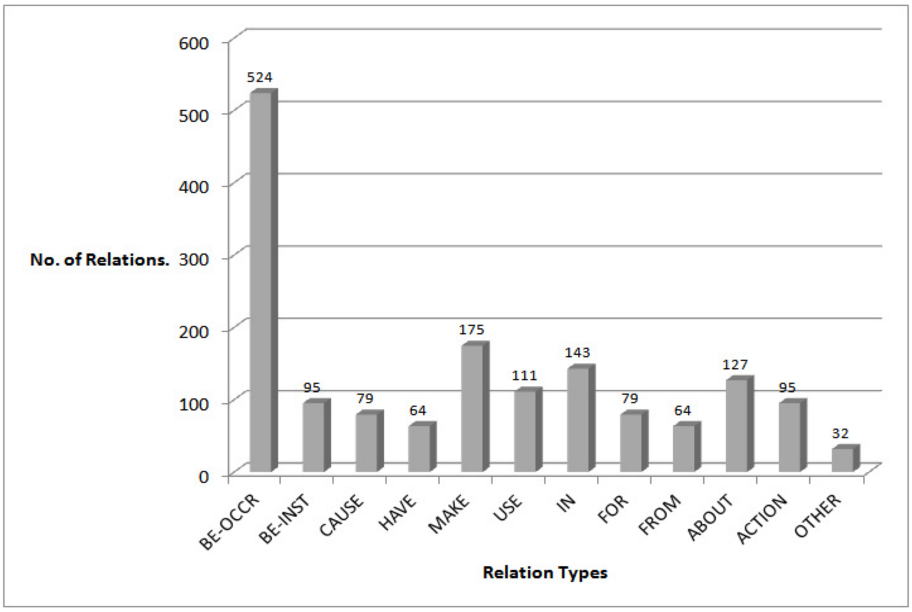

Fig. 1. Figure showing the relative distribution of relation types in a corpus of 20 news paper articles.

6 Discussion

The annotation experiment results strongly indicate that the two natural language usage phenomena of compound noun generation and anaphoric use of nouns are based on the same underlying semantic structure. At a theoretical level, this has a significant impact on our understanding of how humans use natural language. In particular, it will help us better understand the use of compound nouns which are also anaphoric by using the same theory to interpret them. At a computational level, the proposed framework for anaphora resolution allows us to marry the two nlp areas so that we can better share computational advances in the two research areas. Recently there has been an increased momentum [13, 26, 20, 4, 15, 21, 1, 11] towards automatic derivation of relations between composite nouns in noun compounds, most of them based on relations from [18]. This will result in an increasing amount of ontology representing semantic relations used for generating compound nouns. Any such ontology will be directly useful for anaphora resolution in the framework proposed in this paper.

Another significant advantage of a common framework is that it will be easier to integrate the full meaning of a compound noun and the meaning associated with it being used anaphorically. Currently, anaphora is described using a different set of relations (eg. synonymy, hypernomy, meronomy etc.) and compound nouns with a different set. Hence, when interpreting a compound noun which is also anaphoric, it becomes difficult to merge the two meanings. Combining the processes within the same framework gives us a much stronger interpretative power enabling us to interpret a modifier as well as the head noun. As an illustration consider the excerpt below:

> *John's* ***car*** *had an accident yesterday. Its thought* ***faulty car tyres*** *played a major role in the accident.*

The compound noun *faulty car tyres* expresses the relation HAVE between the modifier *car* and the head noun *tyres*, defined by the compound noun generation framework. In terms of straight forward anaphora resolution, the compound noun *faulty car tyres* is not anaphoric since the head noun *tyres* does not co-refer to anything in the previous sentence. However to be able to fully interpret the meaning of the second sentence, it is crucial that we know that the noun *car* in the first sentence also has a HAVE relation to *tyres* in the second sentence. This relation forms the basis of the coherence between "car and accident" in the first sentence and "tyres and accident" in the second sentence. The proposed framework enables us to use relations from the same set to describe the relations between *car* and *tyres* in the compound noun *faulty car tyres* and the anaphoric relations between *faulty car tyres* and *car*. The resultant output from processing a whole discourse using the proposed framework would be a network of entity-to-entity relations consisting of all freely existing nouns as well as nouns participating as modifiers. This network can either be used on its own or used as a building block towards higher level discourse structures such as a coherence structure.

7 Concluding Remarks

In this paper we presented a relational framework for interpreting anaphoric NPs which goes beyond the conventional co-reference relations. We argued that anaphora usage and compound noun generation are based on a common relational framework. To support this we used an existing NP production framework and validated it for anaphora usage using real world discourses. We also argued that by using this framework, a more accurate level of discourse interpretation can be achieved, both directly, as well as using it as a building block for a higher level discourse structure such as the coherence structure. We are in the process of implementing the framework and will be reporting the results in near future. It is anticipated that successful computation of this framework will help in numerous NLP tasks such as document visualization, summarization, archieving/retrieval and search engine applications.

References

1. Barker, K., Szpakowicz, S.: Semi-automatic recognition of noun modifier relationships (1998)

2. Bean, D., Riloff, E.: Corpus-based identification of non-anaphoric noun phrases. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 373–380. Association for Computational Linguistics Morristown, NJ, USA (1999)
3. Bunescu, R.: Associative anaphora resolution: A web-based approach. In: In Proceedings of the EACL2003 Workshop on the Computational Treatment of Anaphora. pp. 47–52 (2003)
4. Butnariu, C., Veale, T.: A concept-centered approach to noun-compound interpretation. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. pp. 81–88. COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), `http://portal.acm.org/citation.cfm?id=1599081.1599092`
5. Davies, M.: Corpus of contemporary american english. http://www.americancorpus.org (2010)
6. Downing, P.: On the creation and use of english compound nouns. Language 53(4), 810–842 (1977)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
8. Fraurud, K.: Definiteness and the processing of noun phrases in natural discourse. Journal of Semantics 7(4), 395–433 (1990)
9. Girju, R.: Improving the interpretation of noun phrases with crosslinguistic information. In: in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 568–575 (2007)
10. Girju, R., Moldovan, D., Tatu, M., Antohe, D.: On the semantics of noun compounds. Computer Speech and Language 19(4), 479–496 (2005)
11. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Séaghdha, D.O., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. SemEval '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), `http://portal.acm.org/citation.cfm?id=1859664.1859670`
12. Hobbs, J.R.: Coherence and coreference. Cognitive Science 67, 67–90 (1979)
13. Kim, S.N., Nakov, P.: Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In: The Proceeding of Conference on Empirical Methods in Natural Language Processing, EMNLP. Edinburgh, UK (2011)
14. Kim, S.N., Baldwin, T.: Automatic interpretation of noun compounds using wordnet similarity. In: In Proceedings of the 2nd International Joint Conference on Natural Language Processing, Jeju Island, South Korea, 1113. pp. 945–956 (2005)
15. Kim, S.N., Baldwin, T.: Interpreting semantic relations in noun compounds via verb semantics. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 491–498. COLING-ACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), `http://portal.acm.org/citation.cfm?id=1273073.1273137`
16. Lauer, M.: Corpus statistics meet the noun compound: some empirical results. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics. pp. 47–54. ACL '95, Association for Computational Linguistics, Stroudsburg, PA, USA (1995), `http://dx.doi.org/10.3115/981658.981665`
17. Levi, J.: Where do all those other adjectives come from. In: Chicago Linguistic Society. vol. 9, pp. 332–354 (1973)
18. Levi, J.N.: The syntax and semantics of complex nominals. Academic Press, New York : (1978)

19. Li, C.N.: Semantics and the Structure of Compounds in Chinese. Ph.D. thesis, University of Carlifornia dissertation. (1971)
20. Nakov, P.: Noun compound interpretation using paraphrasing verbs: Feasibility study (2008), http://www.cs.berkeley.edu/\~{}nakov/selected_papers_list/aimsa2008.pdf
21. Nastase, V.S.S., Sokolova, J., M. Szpakowicz, S.: Learning noun-modifier semantic relations with corpus-based and wordnet-based features. Proceedings of the National Conference on Aritficial Intelligence. 21(Part 1), 781–787 (2006)
22. Nastase, V., Szpakowicz, S.: Exploring noun-modifier semantic relations. In: Proceedings of the 5th International Workshop on Computational Semantics (2003)
23. Ó Séaghdha, D.: Annotating and learning compound noun semantics. In: Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop. pp. 73–78. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://portal.acm.org/citation.cfm?id=1557835.1557852
24. Poesio, M., Vieira, R., Teufel, S.: Resolving bridging references in unrestricted text. In: Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts. pp. 1–6. ANARESOLUTION '97, Association for Computational Linguistics, Morristown, NJ, USA (1997), http://portal.acm.org/citation.cfm?id=1598819.1598820
25. Sanders, T.: Toward a taxonomy of coherence relations. Discourse processes 15(1), 1–35 (1992)
26. Tratz, S., Hovy, E.: A taxonomy, dataset, and classifier for automatic noun compound interpretation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 678–687. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://portal.acm.org/citation.cfm?id=1858681.1858751
27. Vieira, R., Poesio, M.: An empirically based system for processing definite descriptions. Computational Linguistics 26(4), 539–593 (2000), http://www.mitpressjournals.org/doi/abs/10.1162/089120100750105948
28. Warren, B.: Semantic patterns of noun-noun compounds. Gothenburg studies in English, Acta Universitatis thoburgensis (1978)
29. Zimmer, K.E.: Some general observations about nominal compounds. Working Papers on Language Universals pp. C1–21 (1971)