

USING THE WEB TO RESOLVE DEFINITE BRIDGING DESCRIPTIONS

Henry Beecher

September, 2007

Main Reader: Andrew Kehler

Ancillary Reader: Grant Goodall

Ancillary Reader: Ivano Caponigro

Ancillary Reader: Victor Ferreira

Ancillary Reader: Mark Gawron

Ancillary Reader: Rob Malouf

Abstract

A bridging description is a noun phrase whose proper interpretation depends on using general knowledge or inference to identify a previously evoked entity with which it is associated. Conventional computational approaches to anaphora resolution have had little success in overcoming the commonsense barrier posed by definite bridging descriptions. This paper presents preliminary research on a novel alternative involving the adaptation of techniques in automatic knowledge extraction for use with the World Wide Web. Domain independence is achieved without the use of encoded axioms to represent inference mechanisms traditionally assumed by many computational theories. Instead, heuristics are employed which rely only on structural information extracted from corpora, minimal amounts of general hand-coded information, and statistical/frequency information computed from web-search results. These heuristics are combined with techniques like proximity searching and an innovative ‘two-stage’ strategy in a system whose generality and performance can be quantitatively evaluated. The 80% accuracy achieved on targeted bridging descriptions with NP-anchors exceeds both Poesio (2004) and Bunescu (2003), the only two comparable approaches. Plans are also provided for incorporating this initial investigation into a larger dissertation research effort.

USING THE WEB TO RESOLVE DEFINITE BRIDGING DESCRIPTIONS

0. INTRODUCTION.

Bridging Descriptions (Clark, 1977) are a subset of DEFINITE DESCRIPTIONS¹ for which identifying a unique referent requires general knowledge and/or inference about association(s) with a previously evoked entity. Thus in (1), use of the definite description, *the new structure*, is felicitous in so far as the interlocutors (in this case author and reader) both share in the common knowledge that the earlier mentioned ‘G Street Bridge’ refers to an object that maybe classified as a type of structure.

- (1) In Richmond, Ind., the type F railing is being used to replace arched openings on the G Street Bridge. Garret Boone, who teaches art at Earlham College, calls **the new structure** "just an ugly bridge" and one that blocks the view of a new park below.²

Unlike some forms of anaphora resolution in which much can be achieved based almost solely on superficial syntactic cues, the need for a sufficient source of commonsense knowledge remains an obstacle to computational approaches to resolving nominal anaphora with full lexical heads like those of bridging descriptions.

The focus of the research presented here is on overcoming (or circumventing) the commonsense bottleneck in resolving definite bridging descriptions by taking techniques in automatic knowledge extraction previously used on corpora like the British National Corpus (BNC) and adapting them to the largest available corpus, the World-Wide Web. It is only relatively recently that any sort of web-based approaches to anaphora resolution have appeared in the literature, most notably: Bunescu (2003), Poesio, *et al* (2004), and Markert & Nissim (2005). However, while each has reported promising results, their individual efforts have been fairly narrow both in the breadth of data analyzed and in the

¹ Per Russell (1919), any noun phrase beginning with the definite article as in ‘the literature’; thus excluding other definite NPs such as pronouns or possessive descriptions.

² All examples are from the Penn Treebank (1995), unless otherwise noted. Here, underlying and bolding added for purposes of discussion only.

sophistication of techniques used. Nevertheless, these initial attempts provide valuable baselines for gauging the performance of more comprehensive endeavors (an intention explicitly stated by Vieira and Poesio, 2001).

The paper is organized as follows. Essential background information including relevant terminology and the principle factors motivating the proposed research program are provided in §1. A case study on using the Web to resolve bridging descriptions which employs extracts of Wall Street Journal (WSJ) articles found in the Penn Treebank is outlined in §2. The annotation scheme used in constructing training and testing datasets is discussed in §3 and the resulting classification statistics are compared to those reported by other researchers. The heuristics for processing the datasets are detailed in §4. An evaluation of the overall results and error analysis are given in §5 with key similarities and differences to comparable approaches identified. Conclusions and plans for dissertation research are presented in §6.

1. BACKGROUND AND TERMINOLOGY.

1.1 *Classifying Definite Descriptions.*

As shown in (2), bridging anaphora are not restricted to definite descriptions (DDs).

(2) I bought a dictionary today. When I got home, I noticed **a page** had fallen out.³

However, previous research (*cf* Heim, 1982) has found indefinite NPs to be mostly non-anaphoric. Following the approach of M. Poesio and R. Vieira (1998, 2001), in particular, as well as that of: Poesio, *et al* (2004), R. Bunescu (2003), D. Bean and E. Riloff (1999), J. Meyer and R. Dale (2002), I concentrate on bridging anaphora occurring exclusively in DDs. The sizeable literature on DDs comprises the primarily theoretical contributions of Clark (1977), Hawkins (1978), Prince (1981), Heim (1982), and Fraurud (1990), *inter alia* from which Poesio (1998b) has formulated a taxonomy that I also adopt.

Poesio's taxonomy divides DDs into three main categories⁴:

Discourse New: First-mention DD which denotes an object not related by shared associative knowledge to any previously evoked discourse entity; but which includes references to 'larger situation' knowledge (*e.g. the sun, the last century, the White House*), and DDs like *the first man to fly*.

³ Adapted from Prince (1981)

⁴ Poesio & Vieira (1998) reports reliability results for this classification scheme and discuss certain situations in which these categories may not be completely mutually exclusive.

Anaphoric Same-head: Subsequent-mention DD whose resolution simply requires matching the head of the antecedent with that of the definite description as in *a car ... the car*.

Bridging or Inferential: DD whose head is not identical to that of an antecedent, and whose relation with its TRIGGER or ANCHOR⁵ may or may not be one of co-reference. These are a subset of ASSOCIATIVE anaphora in the taxonomy of Hawkins (1978) or of the INFERRABLES class in Prince (1981).

The category of bridging descriptions (BDs) is further sub-divided into those with NP anchors and those without. BDs with NP anchors include:

Synonyms: Anchor and BD are synonymous (and thus co-refer) as in *a new album ... the record*.

Hypernyms/Hyponyms: Anchor and BD are in a *is-a*-relation as in *rice ... the plant* (superordination/hypernymy) or *a plant ... the rice* (subordination/hyponymy). I extend this sub-class to include entities which are examples/members of a category which often occurs with proper nouns such as *Bach ... the composer*, *Zambia ... the African country*, or *AT&T ... the company*.

Meronyms/Holonyms: Anchor and BD are in a *part-of*-relation as in *a tree ... the leaves* (meronymy: the BD is a component of the anchor) or *a bald tire ... the car* (holonymy: the BD has the anchor as a component).

BDs without NP anchors include:

Events: The anchor is introduced by something other than an NP such as a VP or a sentence (e.g. *they planned ... the strategy*).

Discourse Topic: The anchor is an implicit ‘topic’ or ‘theme’ of a discourse, as in *the industry* appearing in a text about oil companies.

General Inference: The anchor is inferred based on more complex relationships such as cause and effect, as in *last week’s earthquake ... the suffering people*.

⁵ Following Hawkins (1978) and Fraurud (1990), these terms indicate a discourse entity with which a bridging description is associated – reserving ‘antecedent’ for identity anaphora.

While these last three BD sub-classes without NP anchors are beyond the scope of the present research effort, any examples are appropriately tagged in all the training and testing data for future use.

1.2 Addressing the Commonsense Bottleneck.

The relationships of synonymy, hyponymy and meronymy cover a very wide range of associations between anchor and BD which interlocutors can accommodate in ordinary language comprehension. Nevertheless, examples like an anchor *U.S. constitution* and BD *the framers* do not clearly belong to any of these three categories⁶. Until recently most proposals⁷ for using computational systems to handle full-NP anaphora, like BDs, have relied on hand-coded resources such as the WordNet or FrameNet ontologies to address the encyclopedic range of lexical and/or world knowledge needed to identify the appropriate anchor. Yet even apart from issues related to building and maintaining such ontologies, Markert and Nissim (2005) point out four major shortcomings to using these resources. First, even large ontologies contain significant knowledge gaps (Vieira and Poesio (2000) report that 62% of the meronymy relations needed for resolving BDs in their corpus were not found in WordNet). Second, context-dependent relations like a reference to *age* as a kind of *factor* are unlikely to be encoded in a fixed, context-independent ontology. Third, word sense proliferation is not systematically encoded so *framer* might likely be associated with *picture* but not *constitution*. Fourth, information encoding is often inconsistent and ideosyncratic so that *door* may be linked to *room* and *house* but not *building*; or *magazine* to *periodic publication* but not to *periodical*.

Researchers like Hearst (1992) and Caraballo (1999) have had only partial success in overcoming some of these deficiencies by extending fixed ontologies through automatic knowledge extraction from corpora. The techniques used for mining corpora typically involve patterns or constructions like ‘NP of NP’ (*e.g.* the office of the president) or ‘NP’s NP’ (*e.g.* the president’s office) to identify instances of meronymy and other relationships. The same sort of lexico-syntactic patterns have since been used by Poesio, *et al* (2002), Markert, Nissim and Modjeska (2003), and Bunescu (2003) not to extend ontologies, but to search large corpora directly and then use the frequency counts generated to gauge the degree of association between a BD and its anchor. The rationale behind this approach stems from the ‘priming’ hypothesis (Poesio *et al*, 1998b)

⁶ In the data reported on here, all BDs were assigned to one of the 3 NP-anchor categories or a catch-all non-NP anchor category. This example was tagged as meronymy.

⁷ *cf* Vieira and Poesio (2000); Harabagiu, Bunescu, and Maiorano (2001); Ng and Cardie (2002b), *inter alia*.

according to which resolving a BD is a matter of finding the anchor in the text which most strongly primes the BD's head predicate. In example (2) above, the text does not explicitly contain *page of a dictionary*; however, high frequency counts from a large corpus for '*page of a dictionary*' or '*dictionary's page*' vs. '*page of a home*' or '*home's page*' is strong evidence that *dictionary* is a more suitable anchor than *home*.

The experience of the researchers who have tried this approach shows that several patterns are required for each relationship category (synonymy, hypernymy, meronymy, *etc.*). Consequently data sparsity is inevitably a problem even with a corpus as large as the BNC containing over 100 million words. For example, the BNC yields no hits for the search string '*page of a dictionary*' and only one hit for '*page of a magazine*'⁸. Furthermore, relying exclusively on priming effects is overly simplistic. Poesio (2003) reports that, in about half of all cases, the actual anchor is not the candidate which is semantically closest to the BD. This often occurs when a particular relation holding between a BD and anchor is situationally defined (*i.e.* context-dependent) as illustrated in (3).

(3) A couple of weeks ago, I lost the case in federal district court in Des Moines. At least, that's the way it was reported. And, indeed, **the lawsuit** was dismissed.

In the context of this example, *the lawsuit* is a synonym of the actual anchor, *case*. However, a competing candidate, *court*, more strongly primes⁹ the BD, *the lawsuit*, than *case* does. Thus, in cases like this, simply choosing the candidate semantically closest to the BD leads to selecting an incorrect anchor. Poesio concludes that BD resolution depends not only on lexical information, but also on other salience factors – either in the sense of being more recent, or of being the 'focus' as suggested by Sidner (1979). Yet, the notoriously difficult task of effectively tracking focus notwithstanding, a preliminary investigation by Poesio (2003) using Centering Theory's parameters (Grosz *et al*, 1995) found less than 60% of anchors to be either the Backward-looking Center (CB) or the Preferred Center (CP).

1.3 Potential for a Web-based Alternative.

Given the shortcomings of fixed ontologies like WordNet, as well as the data-sparsity problems of large structured corpora like the BNC, increasing attention is being paid to treating the Internet as a single massive corpus (indeed the largest available to the

⁸ As compared to 45,900 and 16,000 respectively, using Google.

⁹ As measured by the frequency of *case* versus *court* co-occurring with *lawsuit* adjusted for the overall frequencies of *case* and *court* individually. Cf §4.3.2 and §5.4 for more in-depth discussion.

research community). Using an adequately informed approach, the benefits of the Web's sheer volume and diversity can outweigh its inherent noise and lack of structure (linguistic or otherwise). The Web has been successfully used in several areas of natural language processing (NLP) including machine translation (Grefenstette, 1999) and bigram frequency estimation (Keller *et al*, 2003). Web-based approaches to anaphora resolution have only recently been explored by Bunescu (2003), Poesio, *et al* (2004), and Markert & Nissim (2005).

The efforts of Markert & Nissim (2005) did not include BDs as they focused instead on comparative anaphora (*i.e.* referential NPs with the modifiers *other* or *another* and non-structurally given antecedents). Bunescu (2003), however, reports 53% precision on a corpus containing 324 BDs. The BDs were resolved based on the degree of lexical association between a potential anchor and each BD as measured by calculating a value of pointwise mutual information (PMI)¹⁰. The PMI values were computed using web-search results from permutations of a single phrase pattern " N_t . The N_a *VERB*" where N_t = potential anchor (trigger), N_a = BD head (associate), and *VERB* = one of several generic verbs (*is/are, was/were, has/have, can, etc.*). An example of such a pattern is "...a forest. The trees were...". More recently, Poesio *et al* (2004) report 70.6% precision attempting to resolve specifically mereological (*i.e.* meronymy) BDs in a corpus containing 58 such examples. These BDs were resolved using a multi-layer perceptron (MLP) classifier with back-propagation plus a combination of Google distance, utterance distance, and first-mentioned as features. Google distance was calculated using results from a web query "the *NBD* of the *NPA*" where *NBD* = BD head and *NPA* = potential anchor head. Google distance was 1 if the hit-total = 0, and 1/hit-total otherwise.

With web-based approaches to anaphora resolution still in their nascent stages, I propose conducting a more comprehensive investigation into adapting and refining techniques for taking greater advantage of the World-Wide Web as a resource for resolving bridging descriptions. BDs with NP anchors are the initial focus of this research with the anticipation that progress in this area will improve efforts in resolving BDs without NP anchors as well. The remaining sections describe a preliminary case study into using the Web to resolve BDs involving a sizeable portion of WSJ articles in the Penn Treebank. This study is primarily intended to serve as a proof-of-concept illustrating the merits of the proposal as well as some potential benefits which building on the initial efforts of researchers like Bunescu, Poesio, Vieira, and Markert & Nissim may

¹⁰ *cf.* §4.3.2 for a fuller discussion on PMI and its application in resolving BDs.

achieve. The study also provides an opportunity to verify several results of these earlier efforts, and to re-assess what can be achieved even in the absence of more sophisticated capabilities such as automatic focus tracking.

2. A CASE STUDY: USING THE WEB TO RESOLVE BDs.

2.1 Nature and Sources of Information.

In keeping with earlier research described in §1, a central premise of this approach to resolving BDs is the development of a system whose generality and performance can be quantitatively evaluated – in other words, capable of being tested over a corpus of texts from different domains. Domain independence precludes the use of encoded axioms to represent inference mechanisms traditionally assumed by many computational theories (*cf* Sidner, 1979; Webber, 1979; Carter, 1987; *inter alia*). Instead, the system relies only on structural information extracted from the corpus itself, minimal amounts of general hand-coded information (*e.g.* exception lists), and statistical/frequency information computed from web-search results. Information provided by pre-existing ontological sources like WordNet is not precluded and – although not presently incorporated in this case study – they may have a role in future developments.

WSJ articles in the 1995 version of the Penn Treebank (Marcus, *et al* 1993, 1994) were used for this study because large quantities of these texts are available in a parsed format tagged for parts of speech. A number of the filtering heuristics applied in processing the text are dependent on pre-parsed data. Also, both Vieira and Poesio (2001) as well as Markert and Nissim (2005) use WSJ data from the Penn Treebank; while Bunescu (2003) used Brown data from the Penn Treebank. Thus, critical aspects of the approach described here and its performance can be meaningfully compared to work of these other researchers.

2.2 Important Disclosures regarding Methodology.

To gain sufficient first-hand familiarity with the data and especially the various issues with which any effort to classify DDs is inevitably confronted (*e.g.* split antecedents, co-locations/compounds, proper names, ambiguous uses, *etc*), all the training and testing datasets were newly annotated explicitly for the purposes of this project. Given the number of DDs identified (approximately 2800 in the training dataset and in excess of 1000 in the testing dataset), automating some or all of the annotation task via custom computer code would not have been an unreasonable approach. However, given the exploratory nature of this endeavor, equal consideration had to be given to the downside

of generating, testing, debugging and – especially early on – continually revising custom code. Thus, rather than risk the expenditure of too much time and resources on maintaining code, the entire albeit laborious annotation task was accomplished manually. All other aspects of this project, with the exception of some of the Google search routines, were also carried out manually – that is the filtering, processing heuristics, and result compilations were simulated and/or computed by hand. A margin for error exists whether or not automation is employed – hopefully mitigated here by the fact that all of the data was manually worked through at least twice (initially for annotating and later for processing). A key objective in further extensions of this work, however, is the eventual automation via code of the entire system.

2.3 Case Study Overview.

The major stages of the case study comprised: 1) selecting the corpus; 2) annotating all definite descriptions (DDs) as being one of discourse new, same-head, or bridging anaphora; 3) developing and applying heuristics designed to identify and filter out as many discourse new and same-head examples as possible; 4) treating all remaining examples as bridging descriptions (BDs) and developing additional heuristics designed to identify the best candidate anchor; 5) computing final training-set results and applying the complete process with any refinements to the test data for comparison.

The training dataset used for development and refinement of the system contains 2793 DDs yielding 346 BDs of which 150 have NP anchors (the focus of this study). The test data of approximately 1000 DDs is expected to contain at least 75 BDs with NP anchors. Typically the annotation phase involves 1 or more impartial annotators to avoid potential bias. Being a preliminary proof-of-concept effort, this was not done for the case study; however, it would be an essential aspect of future efforts. Complete details of the annotation scheme adopted and the results of classifying the training data are provided in §3. The heuristics for identifying/segregating discourse new and same-head anaphors as well as the heuristics for selecting the NP anchors of the remaining BDs are described at length in §4. Results from both the training and testing datasets are evaluated in §5 together with an error analysis.

3. THE ANNOTATION SCHEME AND CLASSIFICATION RESULTS.

For the purposes of this case study, annotating the datasets has one ultimate objective: identifying the occurrence of each BD which has a NP anchor. A perfect resolution system would then be able to accomplish 3 things: find the correct anchor for each BD

that has one; identify the BDs without NP anchors; and, identify everything else as non-bridging DDs. The strategy for approaching the ideal system is essentially the same as that used by researchers like Poesio and Bunescu: eliminate the easiest cases to identify and reserve more complex strategies for sorting out the remainder. Thus, labeling all 3 categories of DDs is useful for determining how effectively the system segregates the BDs of interest from the entire set of DDs. All annotations were inserted directly into copies of the tagged and parsed Treebank files.

3.1 *The Discourse New and Same-Head Anaphora Categories.*

Examples of discourse new DDs and same-head anaphora DDs are the most straightforward to annotate. A DD is annotated as new (+N-ew) if it is newly introduced and not in reference to on any previously evoked discourse entity (including an implicit topic). A DD is annotated as same-head anaphora (+I-dentity) if the DD head is a re-occurrence of a previously introduced NP head and it is clearly co-referential. In an example like *a green car* followed later by the DD *the red car*, the DD would not qualify as same-head anaphora. The antecedent of a same-head DD need not itself be a definite description. There is a lack of consensus regarding re-occurring proper nouns which in this project are annotated as discourse new (some researchers treat them as same-head anaphora).

3.2 *The Bridging Category and its Sub-Categories.*

Any DD not annotated as either discourse new or same-head anaphora is considered to be some form of bridging description. The crucial distinction is whether interpreting the BD is dependent on a previously occurring NP anchor. In an example like, *they planned* followed later by the DD *the strategy*, the DD would not qualify as a BD with an NP anchor. The same is typically the case with occurrences of *the matter*, *the problem*, etc. all of which are annotated as BDs without NP anchors (+G-eneral). For the most part making this particular distinction was unproblematic. However, some cases like *the sinking of the Titanic ... the tragedy* could depend on the larger context as to whether or not *the tragedy* was in specific reference to the sinking event itself (*i.e.* hypernymy – *sinking* a kind of *tragedy*). If *the tragedy* was in reference to something only incidentally associated with the *the sinking*, it would more likely be considered a BD without an NP anchor.

Any remaining DDs necessarily do have NP anchors (all other possibilities having already been taken into consideration). The sometimes problematic issue is determining the appropriate relationship between the BD and its anchor. Fortunately, this is not a

crucial distinction in so far as the most immediate objective is resolving all BDs with NP anchors, not just certain varieties¹¹. The usefulness of applying more fine-grained labels lies in being able to measure the effectiveness of various strategies in resolving one type of BD over another when refining the system. Out of 150 BDs with NP anchors more than 98% could justifiably be classified as one of: synonymy (+S), hypernymy (+E), hyponymy (+O), meronymy (+M), or holonymy (+H). A sixth category was not created for the relatively few exceptions (*e.g. the constitution ... the framers*) which were instead subsumed under an existing category – typically meronymy or synonymy. A more common dilemma was distinguishing between synonymy and hypernymy (*e.g. the chemical compound ... the active agent*). The final determination in these cases was often subjective judgment.

3.3 DD Classification Results.

Using the annotation scheme described in §3.1-§3.2, the resulting classification of the 2793 DDs in the training dataset is summarized in Figure (1).

Discourse new:	1896	68%	
Same-head anaphora:	551	19.7%	
Bridging anaphora	346	12.3%	<i>Figure (1)</i>

The results of further sub-classifying the 346 BDs are summarized in Figure (2):

BDs without NP anchors	196	56%	
BDs – synonymy	35	10%	
BDs – hypernymy/hyponymy	102	30%	
BDs – meronymy/holonymy	13	4%	<i>Figure (2)</i>

The outcome of classifying the training dataset is similar to results reported in Vieira and Poesio (2001) in which 1040 DDs¹² comprised 549 (53%) discourse new, 332 (32%) same-head, and 159 (15%) bridging. The subtotal of the discourse new and same-head DDs reported by Vieira and Poesio (2001) is 85% while those two categories constitute 87.7% in this project. The difference in the proportion of same-head anaphora is likely due to how items like proper nouns were treated.

At the outset of this case study there was no way of anticipating or otherwise estimating what the ratio of BDs-with-NP anchors to BDs-without-NP anchors would be.

¹¹ Ultimately, a robust system should be able to identify the relationship holding between a BD and its anchor.

¹² These DDs were extracted from WSJ articles in the Penn Treebank as well.

The few other researchers who have conducted experiments which included quantifying the occurrences of bridging description sub-types in a dataset have each used disparate sub-groupings. In some cases BDs with NP anchors are classified together with BDs whose anchors are a VP or a clause. In other cases, counts for specific types of BDs like that for meronymy or synonymy are the only figures reported. Furthermore, given the more challenging task of identifying non-NP anchors (especially ones associated with a discourse topic or theme), advancing methods for successfully resolving BDs with NP anchors can be an important first step. Understanding the mechanisms by which better methods succeed will likely reveal characteristics about BDs with NP anchors as a group, which in turn can provide essential insights into developing heuristics for differentiating BDs-with-NP anchors from BDs-without-NP anchors.¹³

4. EXPLANATION OF PROCESSING HEURISTICS.

Three sets of heuristics are used in this case study. The first two sets are largely adapted from Vieira and Poesio (2001) and are used to identify and segregate as many non-BDs as possible. The third set of heuristics is used to identify the correct NP anchor for as many BDs as possible. The heuristics in this last set are novel and represent both re-combining some methods used by other researchers and some additional techniques whose application for resolving BDs is not documented elsewhere in the literature.

4.1 *Heuristics for Identifying Discourse New DDs.*

There is no interdependency among these heuristics and they are not listed in any particular order.

- Any DD in the first sentence of a discourse
- DDs which include the following time references (including plural forms and abbreviations): *second, minute, hour, time, day, week, month, year, morning, afternoon, evening, season, past, present, future, age, period, decade, century, span, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, January, February, March, April, May, June, July, August, September, October, November, December, spring, summer, winter, fall*
- DD complements of the copulars *be-seem-become* and their subjects (if DDs)
- DDs that are or contain appositives

¹³*cf* §5 and §6 for further discussion.

- DDs which include: proper nouns; numerics (digits/words); or apostrophe-s possessives
- DDs post-modified by: relative clauses (including null-headed); non-finite clauses (*-ed / -ing* participles and infinitives); or prepositional phrases

The rationale for the last three in particular is essentially the intuition (substantiated in Vieira and Poesio, 2001) that larger or more ‘complex’ DDs are more likely to be discourse new. Put another way, an anaphoric DD is unlikely to require the additional elaboration associated with more complex DDs. Heuristics like these are always a trade off to some degree in so far as they may inadvertently identify some DDs as discourse new which are actually BDs. For example, it is certainly possible for a discourse-initial sentence to contain a BD. The challenge is finding a combination that maximizes the number correctly identified while causing the fewest to be incorrectly identified.

4.2 *Heuristic for Identifying Same-Head Anaphora DDs.*

All same-head anaphora DD are identified through application of the following heuristic.

- Any DD whose head is identical to a preceding NP head provided any modifiers of the DD and those of the NP are not mutually exclusive

This heuristic is intended to avoid identifying as same-head anaphora a combination like *a red car ... the green car*. However, effectively handling such cases goes beyond simply ruling out NPs with mutually exclusive modifiers, as some combinations like *a hybrid vehicle ... the new vehicle* may actually be same-head anaphora. This contrast also draws attention to another dimension of world knowledge needed to properly classify DDs.

4.3 *Heuristics for Identifying the NP Anchors of BDs.*

After application of the heuristics in §4.1-§4.2 any remaining DDs are considered to be bridging descriptions for which an NP anchor needs to be identified. It is important to note at this point that the remaining DDs are in fact still a heterogeneous mixture of BDs with NP anchors, BDs without NP anchors and any discourse new or same-head DDs which the first sets of heuristics failed to identify. For each remaining DD, this last set of heuristics will determine a set of candidate NPs from which the most suitable NP anchor will be selected. Consequently, NP anchors will be inappropriately selected for any DDs which are not actually BDs having NP anchors. Given the results of the initial DD

classification on the training set (*cf* §3.3), minimally 56% are BDs lacking NP anchors.

A NP serving as an anchor for a BD logically precedes it. However, it is both highly inefficient and unnecessary to evaluate every preceding NP in order to determine which one is the appropriate anchor. As found by Sidner (1979), BDs are similar to pronouns in being more sensitive to local focus than other DDs. For this reason researchers often opt to restrict potential anchor candidates to NPs occurring within a prescribed ‘window’. Poesio, *et al* (2002) reports the actual anchors for more than 80% of 204 BDs to occur within a 4 sentence window, and more than 95% within an 8 sentence window. Others choose a specific number of preceding NPs as a window, like Bunescu (2003) who used a window of 50 NPs. Some initial experimentation led to using a window size of 10 NPs for constructing anchor candidate sets in this study. The decision to manually carry out all processing on the data did impose a practical constraint on choosing a window size. However, 10 NPs proved to roughly correlate with a 4-6 sentence window on average, and as discussed later in §5.2, led to misidentifying only 9% of BDs with NP-anchors because the actual anchor was outside of the window.

Preceding NP heads and nominal modifiers each qualify as anchor candidates. Thus in (3), anchor candidates for the BD, *the lawsuit*, include both the NP head, *court*, and the nominal modifier, *district*, found in the preceding NP, *federal district court*. Proper names consisting of two or more nouns qualify both collectively and individually. For example, with *Bank of New England* the entire title serves as an anchor for the BD *the lender*; while just the constituent *New England* could be the anchor for another BD *the region*. For two particular BDs, *the company* and *the firm*, restricting the candidates to proper nouns only (from within the 10-NP window) improved accuracy considerably.

The most suitable anchor for each BD is selected by applying a series of heuristics to members of the BD’s anchor candidate set. Outlined below is an overview of these heuristics which comprise an order-dependent two-stage process (unlike the heuristics described in §4.1-§4.2).

Pre-processing:

- Obtain proximity search counts for each {candidate, BD head} pair from Google

Stage I: Collecting competitors

- Collect as competitors the candidate with the highest value for Pointwise Mutual Information (calculated using the proximity search counts), and any competing candidate(s) in the same order of magnitude as the candidate with the highest PMI

Stage II: Selecting an anchor

- Select from among the collected competitors the one having the single highest hit count across 3 Google searches: “BD or CANDIDATE” (synonymy); “BD like CANDIDATE” (hypernymy); “CANDIDATE’s BD” (meronymy)

Details concerning each of these steps are fully elaborated in the following three sections §4.3.1-§4.3.3.

4.3.1 *Pre-processing Heuristic for Anchor Selection.*

The pre-processing heuristic defined in §4.3 involves using Google to obtain proximity search counts. True proximity searching is a common tool in database retrieval and mining corpora. It is simply defined as retrieving all instances of a given term within a specified word distance of another term. Unfortunately web browser interfaces for Internet search engines like Google, Alta Vista, or Yahoo do not support proximity searching. As an alternative, a Perl script combining the 4 searches listed in Figure (3) was used to emulate proximity searching via the Google application programming interface (API).

“BD CANDIDATE”
 “CANDIDATE BD”
 “BD * CANDIDATE”
 “CANDIDATE * BD”

Figure (3)

In these search strings BD represents the head term of the bridging description being resolved; CANDIDATE represents an NP anchor candidate; and * is a wildcard symbol matching any single word in a Google search¹⁴. The ‘proximity search count’ which the script returns is the aggregate total of the 4 searches. The rationale for using proximity searching as well as its advantages and limitations are discussed in §5.4.

4.3.2 *Stage I Heuristic for Anchor Selection.*

The Stage I heuristic defined in §4.3 involves computing pointwise mutual information (PMI). PMI is an information-theoretic measure representing the degree of relative association between two terms (Manning and Schütze, 1999), in this case the BD head and the head of each anchor candidate. If the anchor candidate is denoted by c and the BD head by b then $PMI(c,b)$ is calculated as shown in Figure (4), where N denotes the total number of documents indexed by Google, $D(query)$ the number of documents returned by Google for the specified query, Q a simple search on a single term, and QI a

¹⁴ The * wildcard only works using the Google API and currently not in the browser interface.

proximity search as described in §4.3.1.

$$\begin{aligned}
 \text{PMI}(c,b) &= \frac{P(c,b)}{\log P(c) * P(b)} \\
 &= \frac{D(QI(c,b)) / N}{\log (D(Q(c)) / N) * (D(Q(b)) / N)} \\
 &= \frac{N * D(QI(c,b))}{\log D(Q(c)) * D(Q(b))}
 \end{aligned}$$

Figure (4)

A PMI value close to 0 indicates virtually no association between an anchor candidate and the BD head. Conversely, higher PMI values indicate a greater degree of relative association between the anchor candidate and the BD head. However, for the purposes of this case study, both N and $D(Q(b))$ are constants across all candidates, so a measure of relative mutual information (RMI) shown in Figure (5) is used instead.

$$\text{RMI}(c,b) = \frac{D(QI(c,b))}{D(Q(c))}$$

Figure (5)

Per the Stage I heuristic, the candidate with the highest PMI (*i.e.* RMI) is selected “and any competing candidate(s) in the same order of magnitude as the candidate with the highest PMI.” This specification is most easily explained with reference to the example in Figure (6) showing actual RMI values calculated for 10 anchor candidates using their respective proximity search counts relative to the BD *the event*.

<u>Candidate</u>	<u>RMI value</u>
<i>building</i>	0.002407
Citicorp	0.000043
model	0.001237
chairs	0.001625
blender	0.000183
gamut	0.000025
<u>show</u>	0.001831
cities	0.001179
Moscow	0.000655
architecture	0.002373

Figure (6)

The correct candidate is show, yet *building* is the candidate with the highest RMI. Given the RMI value of 0.002407 for *building*, any other candidates with RMI values in the range of 0.0002407 to 0.002407 are defined as being in the same order of magnitude as *building*. Since there are other candidates with RMI values in that range (*cf* bold-font

items), *building* is not selected and the resolution process continues on to Stage II.

4.3.3 Stage II Heuristic for Anchor Selection.

The Stage II heuristic defined in §4.3 involves re-processing only candidates whose RMI values are in the same order of magnitude as the candidate with the highest RMI (should there be any). Google hit counts from three exact pattern searches are retrieved for this reduced set of competitors, and the competitor having the single highest count across these searches is selected as the best-suited anchor. These additional pattern searches are designed to identify the single strongest relationship of synonymy, hypernymy, or meronymy among these particular competitors and the BD being resolved. The exact search strings for these pattern searches are listed in Figure (7) below (where BD represents the head of the bridging description being resolved).

Synonymy: “BD or COMPETITOR”

Hypernymy: “BD like COMPETITOR”

Meronymy: “COMPETITOR’s BD”

Figure (7)

To illustrate, Figure (8) shows the results obtained by conducting each of these Google searches using the candidate with the highest RMI in Figure (6) and 5 competitors.

Candidate	“BD or C”	“BD like C”	“C’s BD”
<i>building</i>	16100	126	51
model	5220	8	256
chairs	2	1	735
<u>show</u>	26600	51	546
cities	1	0	3
architecture	1	1	8

Figure (8)

The single highest hit count across all the search results shown in Figure (8) is 26,600. Consequently, the competitor show is selected as the best-suited anchor (which in this case is the correct anchor).

5. EVALUATION OF RESULTS AND ERROR ANALYSIS.

5.1 Evaluation of Results from the Training Dataset.

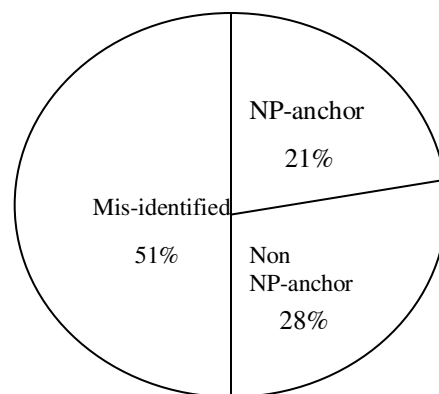
Overall results of processing 2793 DDs in the training dataset are summarized in Figure (9).

Ground Truth	All DDs	Discourse New	Same-head	Bridging	Totals
	Discourse New	1536	0	360	1896
	Same-head	0	545	6	551
	Bridging	0	0	346	346
	Totals	1536	545	712	2793

Figure (9)

The column totals reflect the number of DDs assigned by the system to each of the 3 major categories. The row totals indicate the actual number of DDs in each of these categories as originally annotated (*i.e.* the ground truth¹⁵). The performance of the first set of heuristics in classifying discourse new DDs at 81% accuracy (1536/1896) is in line with results reported by Vieira and Poesio (2001) on their training dataset of 1040 DDs for which discourse new classification achieved an 81% F score¹⁶. The bridging category with 346 BDs includes both BDs with NP anchors and those without, out of which the system correctly resolved 109. The error analysis in §5.2 below provides full details on the resolution accuracy of these 346 BDs as well as on the 6 same-head and 360 discourse new DDs incorrectly classified. A total of 2190 DDs (1536 discourse new, 545 same-head, and 109 bridging) were correctly handled by the system and reflects an overall accuracy of 78.4% (2190/2793). Given the design scope of all the heuristics, this is in keeping with expectations.

More detailed results concerning just the bridging category are summarized in Figure (10).



Ground truth for 712 DDs categorized by system as Bridging

Figure (10)

¹⁵ Ground truth refers to data parameters (here the annotation scheme in §3) used to measure system performance.

¹⁶ The F score used was 2RP/R+P reflecting a combined measure of Recall (correct responses / all cases) and Precision (correct responses / all responses).

This piechart provides a breakdown of the 712 DDs treated by the third set of heuristics as bridging descriptions having NP anchors. As discussed in §4.3, none of the heuristics adapted and/or developed during this case study are designed to classify BDs without NP-anchors. Consequently, 28% of these 712 DDs are the 196 BDs lacking NP anchors. Another 51% are the 366 misidentified discourse new and same-head DDs which are included because, in the present approach, any DDs ‘missed’ by the first two sets of heuristics are automatically assumed to be BDs with NP-anchors. Future development of the system will need to have a more viable alternative to simply lumping together anything that is not previously classified as discourse new or same-head anaphora (*cf* §5.2 below for more discussion). The heuristics developed for this case study are designed to resolve specifically BDs with NP-anchors of which there are 150, constituting 21% of the 712. This group is also 43% (150/346) of all actual bridging descriptions, both with and without NP-anchors.

Specific results concerning only the BDs with NP-anchors are summarized in Figure (11).

BD with NP-anchors	Actual	Correct	Incorrect
Synonymy	35 (23%)	23	12
Hypernymy	102 (68%)	74	28
Meronymy	13 (9%)	12	1
Totals	150	109 (73%)	41 (27%)

Figure (11)

This table depicts an analysis of the 150 BDs with NP-anchors and provides a breakdown according to the type of relation holding between the BD and its anchor¹⁷. It is interesting that while meronymy represented the smallest subset of these BDs, this is also the group on which the heuristics perform best. The 73% accuracy in resolving BDs with

¹⁷ This breakdown by relation is provided for discussion purposes only. The heuristics developed thus far do not differentiate among types of relations holding between a BD and its anchor in all cases.

NP-anchors is the most relevant measure of the performance of the heuristics on the target DDs for which they are designed. With an accuracy of 73% the performance of these novel heuristics in resolving specifically BDs with NP-anchors exceeds that of comparable approaches reported in the literature to date (*cf* §1.3 and §5.4). The 41 BDs incorrectly resolved are discussed in the error analysis given in §5.2 below. However, it is relevant to note that in 14 of 41 incorrectly resolved cases, the actual anchor is outside of the 10-word window used to generate anchor candidate sets. Thus the 109 correctly resolved BDs represent 80% of 136 BDs with NP-anchors occurring within the scope of the heuristics.

5.2 Error Analysis on Training Dataset.

In the entire training dataset 603 DDs (21.6%) proved to be problematic for the heuristics as designed for this case study. The largest group (60% of the 603) consisted of 360 DDs which the system failed to identify as discourse new. These 360 DDs also represent 19% of the total 1896 discourse new DDs in the training dataset. The same proportion in this category is reported as misclassified by Vieira and Poesio (2001) from whom the heuristics for identifying them were adapted. As previously discussed (*cf* §4.1) these heuristics largely exploit the strong tendency for more complex DDs to be discourse new. Consequently, simple often one-word DDs like *the sun*, *the Iraq War*, or *the nation* remain undetected as discourse new after being processed by the first set of heuristics. The majority of these DDs are described as ‘larger situation’ uses in the taxonomy of Hawkins (1978) because they refer to an entity or event whose existence is of common knowledge. In addition, many DDs appearing in idioms such as *the bucket* in *kick the bucket* also go unidentified as discourse new using these heuristics.

The system also failed to identify 6 DDs as same-head anaphora. Representing less than 0.5% of the 603 problematic DDs, these are all same-head anaphora which failed the condition barring ‘mutually exclusive modifiers’ in the heuristic for identifying them (*cf* §4.2). In other words, the same-head DD and antecedent each had a modifier the other did not.

Another 32.5% of the 603 problematic DDs are the 196 BD without NP-anchors. Although these are bridging descriptions, they are incorrectly resolved with NP anchors because the system is not presently equipped to distinguish them from BDs which do have NP-anchors. Experience thus far with designing heuristics to resolve the BDs with NP-anchors has shown the effectiveness of strategies exploiting RMI values from proximity searches used to measure lexical association (*cf* §5.4 for further discussion and details). Similar strategies could be extended to differentiate BDs without NP-anchors,

and possibly the ‘larger situation’ type discourse new DDs presently escaping detection. A careful examination and analysis of RMI values and other lexical association measures obtained in connection with BDs lacking NP-anchors and larger-situation DDs may reveal patterns and/or thresholds useful for correctly identifying them. Efforts along these lines were beyond the scope of this case study, but are a high priority in future development work on the system.

The final group of 41 (7% of the 603 problematic DDs) are BDs with NP-anchors which were resolved with the wrong NP. As indicated in §5.1, the correct anchor was unattainable in 14 of these cases (9% of the total 150 BDs with BP-anchors) because it was outside of the 10-word window used to generate anchor candidate sets. In another 3 cases the BDs had split anchors (like the example in (4) below) and the system failed to identify at least one of them.¹⁸

(4) The move is designed to ward off a hostile takeover attempt by two European shipping concerns, Stena Holding AG and Tiphook PLC. In May, **the two companies**, through their jointly owned holding company, Temple, offered \$50 a share, or \$777 million, for Sea Containers.

Some of the remaining 24 cases fell into one of two diametrically opposed circumstances. On the one hand, in examples like *the university president* having *Mr. Hahn* as an anchor, or *the beverage carrier* having *cup-tote* as an anchor, the relationships between the BDs and anchors are sufficiently obscure to make the search values used for measuring lexical association ineffective in identifying the correct anchor. On the other hand, in examples like *the thing* having *product* as an anchor, or *the man* having *a mainland Chinese* as an anchor, the BDs are sufficiently generic to also make measures of lexical association ineffective in identifying the correct anchor. In addition there are cases in which the heuristics are confronted with two or more NPs in an anchor candidate set that are essentially interchangeable with respect to a particular BD, even though only one is actually the anchor. Examples of this include the candidate set for *the company* containing both *HP* and *IBM*; or the candidate set for *the country* containing both *France* and *Germany*. The performance of the heuristics on some of these cases may be further improved by incorporating additional factors such as giving more weight to candidates functioning in a subject role, or to candidates whose word distance is

¹⁸ For the purposes of this study, BDs with split anchors were considered correctly resolved if at least one of the NPs was identified. A more robust system would need to correctly identify all the NPs involved.

closest to the BD. Future development work on the system will examine the effectiveness of including these and other related strategies.

5.3 Evaluation of Results from the Testing Dataset.

The course of adapting and developing the current version of the heuristics involved re-processing portions of the training data multiple times. Consequently there is a tendency for aspects of the heuristics to over-conform to the dataset used in development. To assure bias-neutral results, standard practice is to set aside a portion of the data to be used only in final testing for comparison purposes. The performance of the system on the test data in relation to its performance on the training data is reviewed in this section.

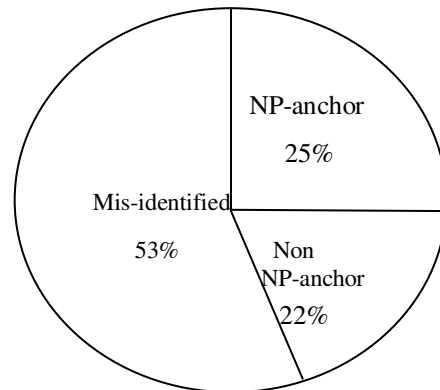
Overall results of processing the 1033 DDs in the testing dataset are summarized in Figure (12).

	All DDs	Discourse New	Same-head	Bridging	Totals
Ground Truth	Discourse New	529	0	136	665
	Same-head	0	201	25	226
	Bridging	2	0	140	142
	Totals	531	201	301	1033

Figure (12)

Performance on classifying specifically discourse new DDs at 80% accuracy (529/665) is in keeping with results from the training dataset which achieved 81% accuracy in discourse new classification. The 25 same-head anaphora DDs incorrectly classified are 11% of the total in this category and a significantly higher proportion than the 0.5% misclassified in the training dataset. In all likelihood this discrepancy is mostly the result of better detecting DDs in this group which fail the condition barring ‘mutually exclusive modifiers’ (*cf* §4.2). The two bridging DDs misclassified were both time references which the heuristics uniformly treat as discourse new. The system correctly handled a total of 779 DDs in the testing dataset (529 discourse new, 201 same-head, and 49 bridging) reflecting an overall accuracy of 75% (779/1033) as compared to 78.4% overall accuracy in the training dataset.

More detailed results on the bridging category alone are summarized in Figure (13).



Ground truth for 301 DDs categorized by system as Bridging

Figure (13)

This piechart breaks down the 301 DDs treated by the third set of heuristics as bridging descriptions having NP anchors. The overall ratios are very close to those of the training data with slightly more misidentified (161/301 or 53% in the testing data as opposed to 51% in the training data). The largest difference is with BDs lacking NP-anchors of which there were 67 out 301 DDs or 22% versus 28% in the training data. Conversely 75 of the 301 DDs were BDs with NP-anchors constituting 25% as opposed to 21% for this group in the training data. All the actual BDs (with and without NP-anchors) in the testing data totaled 142 or 14% (142/1033) versus 12% in training data. The testing data contained 75 actual BDs with NP-anchors or 53% (75/142) of all actual BDs in contrast to 43% (150/346) in the training data.

Specific results concerning only the BDs with NP-anchors are summarized in Figure (14).

BD with NP-anchors	Actual	Correct	Incorrect
Synonymy	16 (21%)	9	7
Hypernymy	50 (67%)	36	14
Meronymy	9 (12%)	4	5
Totals	75	49 (65%)	26 (35%)

Figure (14)

For the 75 BDs with NP-anchors in the testing data, the breakdown according to type of relation between a BD and its anchor is very similar to that in training data (which had 23% synonymy; 68% hypernymy; 9% meronymy). An accuracy of 65% on these BDs is less than the 73% achieved on the training data. However, in the testing data, the percentage of incorrectly resolved cases with actual anchors outside the 10-word window used to generate anchor candidate sets is 38% (10/26). This is a larger proportion than the 34% (14/41) in the training data. Thus, the 49 correctly resolved BDs represent 75% of the 65 BDs with NP-anchors occurring within the scope of the heuristics. While below the 80% (109/136) in the training data for BDs with NP-anchors in the scope of the heuristics, 75% still exceeds comparable approaches reported in the literature to date (*cf* §1.3 and §5.4).

5.4 Similarities and Differences with Comparable Approaches.

As previously pointed out, the basic strategy of isolating BDs by first sorting out discourse new and same-head anaphora DDs is adopted from Vieira and Poesio (2001) and was used by Bunescu (2003) as well. Poesio, *et al* (2002), Bunescu (2003), and Markert, Nissim and Modjeska (2003) all make use of pointwise mutual information and particular pattern searches although none include these tactics in the same way as found in the heuristics developed for this case study. Bunescu (2003) and Poesio (2004) both specifically target BDs with NP-anchors, thus replicating either approach using the training data from the present case study would provide a meaningful comparison. Unfortunately, Bunescu used a window of 50 NPs which cannot be feasibly replicated in the absence of automated procedures. Poesio, on the other hand, employed a customized MLP classifier, the details for which were not published.

One key difference with the heuristics used in this study and each of these cases is their reliance on PMI values calculated from only a single search pattern. As discussed in §1.3, in some cases the single pattern attempted to accommodate any possible relation between a BD and anchor (*e.g.* “ N_t . The N_a VERB” used by Bunescu, 2003). In other cases, a single pattern was intended for a specific relation like meronymy (*e.g.* “the NBD of the NPA” used by Poesio, 2004). Furthermore, the use of only a single-stage process in these other approaches is also a key difference. Employing proximity searching and a two-stage process are both novel aspects which differentiate the approach used here from any other found in the literature.

The idea to incorporate proximity searching developed out of a number of failed attempts to gauge the strength of association between a BD head and its anchor using only permutations of various pattern searches. While one specific pattern may be very

effective for a particular kind of relationship (*i.e.* synonymy, hypernymy, meronymy), there is really no single pattern or set of patterns that is equally effective across all the most common relationships. Another option is testing the entire set of candidates for all three relationships, but this too has major limitations. Not infrequently the actual relationship of a BD to its anchor may be one type, like meronymy; but another type, like synonymy, is co-incidentally more strongly shared with some other candidate. Just as problematic is finding two or more candidates for which the pattern search results are in such close range to each other that a clear preference is indiscernible with any amount of confidence.

In reality these problems cannot be completely eliminated (at least not with resources and capabilities presently available). However, more sophisticated strategies can help minimize factors preventing the appropriate candidate from being determined with greater confidence. Rather than initially using specific tests (*i.e.* patterns) for specific relations, a single more general indication of each candidate's overall degree of association with a BD could be effective in at least separating more likely candidates from less likely ones. Proximity searching is ideally suited to this. A proximity search for a candidate within a 1-word distance (in either direction) of a BD effectively subsumes many relation-type specific patterns such as "C or BD"; "BD or C"; "C of BD"; "BD's C"; "BD like C" as well as collocations (0-word distance) like "C BD" or "BD C". On the other hand, this kind of searching is also a rather weak filter because there is no restriction on intervening words resulting in large amounts of noise.¹⁹ Nevertheless, it is still a much more broad-based measure of lexical association. In a pilot test of 50 BDs (all with NP anchors), selecting a candidate based on the results of proximity searching alone achieved an encouraging 50% accuracy.

Not surprisingly, relying solely on the 'raw' results of proximity searching gives some candidates, in effect, an unfair advantage simply because they occur much more frequently than other candidates. This advantage is eliminated by adjusting the 'raw' results relative to the degree of each candidate's individual frequency – which is exactly what calculating relative mutual information (RMI) accomplishes. Using RMI to re-evaluate the proximity search results in the pilot of 50 BDs boosted accuracy to 70% (a 40% increase). Even at this shallow level of processing, these results are not far off from what others have achieved. As discussed in §1.3, Bunescu (2003) reports 53% precision in resolving BDs through calculating PMI values on permutations of a single pattern " N_t . The N_a VERB". While Poesio, *et al* (2004) report 70.6% precision resolving specifically

¹⁹ The term 'noise' refers to either undesirable results or any possible garbage, gibberish, *etc.*

meronymy-type BDs using a multi-layer perceptron (MLP) classifier.

Further analyzing the RMI values obtained in the pilot of 50 BDs revealed an interesting ‘clustering’ effect. In approximately two-thirds of the correctly identified cases, the highest RMI value exceeded its closest competitor by more than an order of magnitude. Conversely, this held true in less than one-third of the misidentified cases. This inverse clustering indicates that a best-suited anchor cannot be reliably singled out in a majority of the misidentified cases because candidates have RMI values too close to one another. Thus there is good reason in such cases to re-evaluate any candidates whose RMI values are in close range to the highest RMI value. How to best re-evaluate these cases, though, is an open question. On a broad-based measure of association with the BD, there is little distinguishing members of the reduce candidate set. So, at stage II, it could now be more effective to test for all 3 relation types (via the 3 pattern searches in §4.3.3); and then to select the candidate with the single overall highest count as the best-suited anchor. After taking this approach to re-evaluate 20 of the 50 pilot BDs whose highest-RMI candidate had competitors in the same order of magnitude, net accuracy further increased to 84% (42/50 reflecting a gain of 10 previously misidentified, and a loss of 3 correctly identified). These initial results provided sufficient corroboration to justify further testing the approach on the full training dataset.

One additional observation is worth noting. At least superficially, the approach to BD resolution developed in this case study has some interesting parallels to theoretical ideas presented in Sidner (1979). She proposed that anaphors including bridging references are resolved by a two-stage process in which initial hypotheses prompted on the basis of focusing information are subsequently accepted or rejected through commonsense inference. Calculating RMI values based on proximity searching in the Stage I heuristics appears to identify the most salient candidate(s) with respect to the BD. Then (if needed) picking the one with the strongest relationship of synonymy, hypernymy or meronymy among these salient candidates, the Stage II heuristics increase the odds of zeroing in on the correct anchor.

6. CONCLUSION

6.1 *Key Findings of the Case Study.*

Among the most significant findings were the results from resolving specifically BDs with NP anchors. Achieving 80% accuracy in the training dataset on such cases within the scope of the heuristics, and 75% accuracy on those in test dataset, exceed any comparable results yet reported in the literature. This level of performance is an

empirical demonstration of the viability of combating the commonsense bottleneck in resolving bridging anaphora by adapting to the web more innovative techniques in corpus analysis and other computational methods. In addition this study provides a valuable source of comparison to other nascent research in using comparable web-based approaches. The training dataset's ratio of 12% BDs to 88% (combined) discourse new and same-head anaphora DDs, as well as the test dataset's ratio of 14% to 86%, support the findings of Vieira and Poesio (2001) who report a 15% to 85% ratio. This correlation is an indication that the taxonomy Poesio has adapted from Clark (1977) and Hawkins (1978) *inter alia* is largely adequate at the highest DD divisions (*i.e.* discourse new, same-head, and bridging). However, there is less consensus with regards to more fine-grained sub-divisions. Some researchers (*e.g.* Bunescu, 2003) do not distinguish BDs with NP anchors from a larger set of BDs whose anchors are any explicit discourse entities (including VPs and clauses) versus BDs whose anchors are not explicit. Furthermore, the types of relations holding between BD and anchor are not limited to synonymy, hypernymy and meronymy; although just what additional categories to include and especially how to most appropriately classify borderline cases remain unresolved issues.

Equally valuable among the case study's findings are some of its deficiencies. Up to 20% of the discourse-new DDs in both the training and test datasets eluded detection by the first set of heuristics. Current research into such properties of definiteness as familiarity and uniqueness (*cf* Lyons, 1999) could provide further insight into distinguishing between anaphoric and discourse-new uses of 'simple' DDs (*i.e.* those without complex modifiers like relative clauses). For instance, relational DDs like *the mother* or *the head* while often anaphoric are not exclusively so. Thus, a computational means of determining when they are (or are not) could contribute greatly to reducing the number of discourse-new DDs escaping detection. Another major limitation is the lack of any heuristics for differentiating BDs with NP-anchors from BDs without NP-anchors. This is an issue which cannot go unaddressed given that 56% of all BDs in the training dataset are without NP-anchors, and 47% in the testing dataset. Better understanding the properties and behavior of BDs with NP-anchors (from a computational perspective) should contribute to efforts in differentiating them from BDs without NP-anchors. Here too, theoretical research on the properties of definiteness may provide some valuable insights as well. For example, DDs like *the problem*, *the issue* or *the matter* frequently, but not exclusively, function as BDs without NP-anchors. Consequently, being able to reliably ascertain when they do (or do not) would greatly enhance the development of new heuristics.

REFERENCES

- Bean, David and Ellen Riloff (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL-99), pp. 373-380. Providence, RI.
- Bunescu, Razvan (2003). Associative anaphora resolution: A web-based approach. In R. Dale, K. van Deemter, and R. Mitkov, editors, *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*.
- Byron, Donna (2004). *Resolving Pronominal Reference to Abstract Entities*. Ph.D. thesis, University of Rochester, Dept. of Computer Science, Rochester, NY.
- Caraballo, Sharon (1999). Automatic acquisition of a hypernym-labelled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL-99), pp. 120-126. Providence, RI.
- Carter, David (1987). *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.
- Clark, Herbert (1977). Inferences in comprehension. In D. Laberge and S. Samuels, editors, *Basic Process in Reading: Perception and Comprehension*. Lawrence Erlbaum, pp. 243-263.
- Cruse, David (1986). *Lexical Semantics*. Cambridge University Press: Cambridge, UK.
- Fraurud, Keri (1990). Definiteness and the processing of NPs in natural discourse. In *Journal of Semantics*, v7, pp. 395-433.
- Grefenstette, Gregory (1999). The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99: Translating and the Computer 21*, London, UK.
- Grosz, Barbara, J. Aravind, K. Joshi and S. Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. In *Computational Linguistics*, v21:2, pp.202-225.
- Harabagiu, Sanda, Razvan Bunescu and Steven Maiorano (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the Second Conference of the North American Chapter of the ACL*, pp. 55-62. Pittsburgh.
- Hawkins, John (1978). *Definiteness and Indefiniteness*. Croom Helm, London, UK.
- Hearst, Marti (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France.
- Heim, Irene (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst, MA.

- Keller, Frank and Maria Lapata (2003). Using the Web to obtain frequencies for unseen bigrams. In *Computational Linguistics*, v29:3, pp. 459-484.
- Lyons, Christopher (1999). *Definiteness*. Cambridge University Press: Cambridge, UK.
- Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, v19:2, pp. 313-330.
- Markert Katja and Malvina Nissim (2005). Comparing knowledge sources for nominal anaphora resolution. In *Computational Linguistics*, v31:3, pp.367-402.
- Markert, Katja, Malvina Nissim and Natalia Modjeska (2003). Using the Web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pp. 39-46, Budapest.
- Meyer, Josef and Robert Dale (2002). Mining a corpus to support associative anaphora resolution. In *Proceedings of the Fourth International Conference on Discourse Anaphora and Anaphor Resolution*, Lisbon.
- Ng, Vincent and Claire Cardie (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104-111, Philadelphia, PA.
- Poesio, Massimo, Sabine Schulte im Walde and Chris Brew (1998b). Lexical clustering and definite description interpretation. In *Proceedings of the AAAI Spring Symposium on Learning for Discourse*, pp. 82-89, Stanford, CA.
- Poesio, Massimo and Renata Vieira (1998). A corpus-based investigation of definite description use. In *Computational Linguistics*, v24:2, pp. 183-216.
- Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde and Renata Vieira (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 1220-12224, Las Palmas, Canary Islands.
- Poesio, Massimo (2003). Associative descriptions and salience. In *Proceedings of the EACL Workshop on Computational Treatments of Anaphora*.
- Poesio, Massimo, Rahul Mehta, Axel Maroudas and Janet Hitzeman (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 143-150, Barcelona.
- Prince, Ellen (1981). Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*. pp. 223-256. Academic Press: New York.

- Russell, Bertrand (1919). Descriptions. In *Introduction to Mathematical Philosophy*. George Allen and Unwin Publishers.
- Sidner, Candace (1979). *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT, MA.
- Vieira, Renata and Massimo Poesio (2000b). Processing definite descriptions in corpora. In S. Botley and T. McEnery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*. pp. 189-212. John Benjamins: Amsterdam.
- Vieira, Renata and Massimo Poesio (2001). An empirically-based system for processing definite descriptions. In *Computational Linguistics*, v26:4.
- Webber, Bonnie (1979). *A Formal Approach to Discourse Anaphora*. Garland: New York.