Resolving "This-issue" Anaphora

Varada Kolhatkar
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
varada@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
gh@cs.toronto.edu

Abstract

We annotate and resolve a particular case of abstract anaphora, namely, *this-issue* anaphora. We propose a candidate ranking model for *this-issue* anaphora resolution that explores different *issue*-specific and general abstract-anaphora features. The model is not restricted to nominal or verbal antecedents; rather, it is able to identify antecedents that are arbitrary spans of text. Our results show that (a) the model outperforms the strong adjacent-sentence baseline; (b) general abstract-anaphora features, as distinguished from *issue*-specific features, play a crucial role in *this-issue* anaphora resolution, suggesting that our approach can be generalized for other NPs such as *this problem* and *this debate*; and (c) it is possible to reduce the search space in order to improve performance.

1 Introduction

Anaphora in which the anaphoric expression refers to an abstract object such as a proposition, a property, or a fact is known as *abstract object anaphora*. This is seen in the following examples.

(1) [**Be careful what you wish... because wishes sometimes come true.**]$_i$ [**That**]$_i$ is what the Semiconductor Industry Association, which represents U.S. manufacturers, has been learning. (from Asher (1993))

(2) This prospective study suggested [**that oral carvedilol is more effective than oral metoprolol in the prevention of AF after on-pump CABG**]$_i$. It is well tolerated when started before and continued after the surgery. However, further prospective studies are needed to clarify [**this issue**]$_i$.

(3) In principle, he said, airlines should be allowed **[to sell standing-room-only tickets for adults]**$_i$ — as long as **[this decision]**$_i$ was approved by their marketing departments.

These examples highlight a difficulty not found with nominal anaphora. First, the anaphors refer to abstract concepts that can be expressed with different syntactic shapes which are usually not nominals. The anaphor *That* in (1) refers to the proposition in the previous utterance, whereas the anaphor *this issue* in (2) refers to a clause from the previous text. In (3), the anaphoric expression *this decision* refers to a verb phrase from the same sentence. Second, the antecedents do not always have precisely defined boundaries. In (2), for example, the whole sentence containing the marked clause could also be thought to be the correct antecedent. Third, the actual referents are not always the precise textual antecedents. The actual referent in (2), the issue to be clarified, is *whether oral carvedilol is more effective than oral metoprolol in the prevention of AF after on-pump CABG or not*, a variant of the antecedent text.

Generally, abstract anaphora, as distinguished from *nominal anaphora*, is signalled in English by pronouns *this*, *that*, and *it* (Müller, 2008). But in abstract anaphora, English prefers demonstratives to personal pronouns and definite articles (Passonneau, 1989; Navarretta, 2011).[1] Demonstra-

[1] This is not to say that personal pronouns and definite articles do not occur in abstract anaphora, but they are not common.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1255–1265, Jeju Island, Korea, 12–14 July 2012.

tives can be used in isolation (*That* in (1)) or with *nouns* (e.g., *this issue* in (2)). The latter follows the pattern *demonstrative* $\{modifier\}*$ *noun*. The demonstrative acts as a determiner and the noun following the demonstrative imposes selectional constraints for the antecedent, as in examples (2) and (3). Francis (1994) calls such nouns *label nouns*, which "serve to encapsulate or package a stretch of discourse". Schmid (2000) refers to them as *shell nouns*, a metaphoric term which reflects different functions of these nouns such as encapsulation, pointing, and signalling.

Demonstrative nouns, along with pronouns like *both* and *either*, are referred to as *sortal anaphors* (Castaño et al., 2002; Lin and Liang, 2004; Torii and Vijay-Shanker, 2007). Castaño et al. observed that sortal anaphors are prevalent in the biomedical literature. They noted that among 100 distinct anaphors derived from a corpus of 70 Medline abstracts, 60% were sortal anaphors. But how often do demonstrative nouns refer to abstract objects? We observed that from a corpus of 74,000 randomly chosen Medline[2] abstracts, of the first 150 most frequently occurring distinct demonstrative nouns (frequency > 30), 51.3% were abstract, 41.3% were concrete, and 7.3% were discourse deictic. This shows that abstract anaphora resolution is an important component of general anaphora resolution in the biomedical domain. However, automatic resolution of this type of anaphora has not attracted much attention and the previous work for this task is limited.

The present work is a step towards resolving abstract anaphora in written text. In particular, we choose the interesting abstract concept *issue* and demonstrate the complexities of resolving *this-issue* anaphora manually as well as automatically in the Medline domain. We present our algorithm, results, and error analysis for *this-issue* anaphora resolution.

The abstract concept *issue* was chosen for the following reasons. First, it occurs frequently in all kinds of text from newspaper articles to novels to scientific articles. There are 13,489 *issue* anaphora instances in the New York Times corpus and 1,116 instances in 65,000 Medline abstracts. Second, it is abstract enough that it can take several syntactic and semantic forms, which makes the problem interesting and non-trivial. Third, *issue* referents in scientific literature generally lie in the previous sentence or two, which makes the problem tractable. Fourth, *issues* in Medline abstracts are generally associated with clinical problems in the medical domain and spell out the motivation of the research presented in the article. So extraction of this information would be useful in any biomedical information retrieval system.

[2] http://www.nlm.nih.gov/bsd/pmresources.html

2 Related Work

Anaphora resolution has been extensively studied in computational linguistics (Hirst, 1981; Mitkov, 2002; Poesio et al., 2011). But CL research has mostly focused on nominal anaphora resolution (e.g., resolving multiple ambiguous mentions of a single entity representing a person, a location, or an organization) mainly for two reasons. First, nominal anaphora is the most frequently occurring anaphora in most domains, and second, there is a substantial amount of annotated data available for this kind of anaphora.

Besides pronominal anaphora, some work has been done on complement anaphora (Modjeska, 2003) (e.g., *British and **other** European steelmakers*). There is also some research on resolving sortal anaphora in the medical domain using domain knowledge (Castaño et al., 2002; Lin and Liang, 2004; Torii and Vijay-Shanker, 2007). But all these approaches focus only on the anaphors with nominal antecedents.

By contrast, the area of abstract object anaphora remains relatively unexplored mainly because the standard anaphora resolution features such as agreement and apposition cannot be applied to abstract anaphora resolution. Asher (1993) built a theoretical framework to resolve abstract anaphora. He divided discourse abstract anaphora into three broad categories: event anaphora, proposition anaphora, and fact anaphora, and discussed how abstract entities can be resolved using discourse representation theory. Chen et al. (2011) focused on a subset of event anaphora and resolved event coreference chains in terms of the representative verbs of the events from the OntoNotes corpus. Our task differs from their work as follows. Chen et al. mainly

focus on events and actions and use verbs as a proxy for the non-nominal antecedents. But *this-issue* antecedents cannot usually be represented by a verb. Our work is not restricted to a particular syntactic type of the antecedent; rather we provide the flexibility of marking arbitrary spans of text as antecedents.

There are also some prominent approaches to abstract anaphora resolution in the spoken dialogue domain (Eckert and Strube, 2000; Byron, 2004; Müller, 2008). These approaches go beyond nominal antecedents; however, they are restricted to spoken dialogues in specific domains and need serious adaptation if one wants to apply them to arbitrary text.

In addition to research on resolution, there is also some work on effective annotation of abstract anaphora (Strube and Müller, 2003; Botley, 2006; Poesio and Artstein, 2008; Dipper and Zinsmeister, 2011). However, to the best of our knowledge, there is currently no English corpus annotated for *issue* anaphora antecedents.

3 Data and Annotation

To create an initial annotated dataset, we collected 188 *this* {*modifier*}* *issue* instances along with the surrounding context from Medline abstracts.[3] Five instances were discarded as they had an unrelated (publication related) sense. Among the remaining 183 instances, 132 instances were independently annotated by two annotators, a domain expert and a non-expert, and the remaining 51 instances were annotated only by the domain expert. We use the former instances for training and the latter instances (unseen by the developer) for testing. The annotator's task was to mark arbitrary text segments as antecedents (without concern for their linguistic types). To make the task tractable, we assumed that an antecedent does not span multiple sentences but lies in a single sentence (since we are dealing with singular *this-issue* anaphors) and that it is a continuous span of text.

[3]Although our dataset is rather small, its size is similar to other available abstract anaphora corpora in English: 154 instances in Eckert and Strube (2000), 69 instances in Byron (2003), 462 instances annotated by only one annotator in Botley (2006), and 455 instances restricted to those which have only nominal or clausal antecedents in Poesio and Artstein (2008).

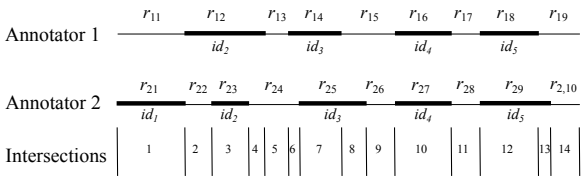


Figure 1: Example of annotated data. Bold segments denote the marked antecedents for the corresponding anaphor *ids*. r_{hj} is the j^{th} section identified by the annotator h.

3.1 Inter-annotator Agreement

This kind of annotation — identifying and marking arbitrary units of text that are not necessarily constituents — requires a non-trivial variant of the usual inter-annotator agreement measures. We use Krippendorff's *reliability coefficient for unitizing* (α_u) (Krippendorff, 1995) which has not often been used or described in CL. In our context, *unitizing* means marking the spans of the text that serve as the antecedent for the given anaphors within the given text. The coefficient α_u assumes that the annotated sections do not overlap in a single annotator's output and our data satisfies this criterion.[4] The general form of coefficient α_u is:

$$\alpha_u = 1 - \frac{{}_uD_o}{{}_uD_e} \quad (1)$$

where ${}_uD_o$ and ${}_uD_e$ are observed and expected disagreements respectively. Both disagreement quantities express the average squared differences between the mismatching pairs of values assigned by annotators to given units of analysis. $\alpha_u = 1$ indicates perfect reliability and $\alpha_u = 0$ indicates the absence of reliability. When $\alpha_u < 0$, the disagreement is systematic. Annotated data with reliability of $\alpha_u \geq 0.80$ is considered reliable (Krippendorff, 2004).

Krippendorff's α_u is non-trivial, and explaining it in detail would take too much space, but the general idea, in our context, is as follows. The annotators mark the antecedents corresponding to each anaphor in their respective copies of the text, as shown in Figure 1. The marked antecedents are mutually exclusive sections r; we denote the j^{th} section identified

[4]If antecedents overlap with each other in a single annotator's output (which is a rare event) we construct data that satisfies the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.

Antecedent type	Distribution	Example
clause	37.9%	There is a controversial debate (**SBAR** *whether back school program might improve quality of life in back pain patients*). This study aimed to address **this issue**.
sentence	26.5%	(**S** *Reduced serotonin function and abnormalities in the hypothalamic-pituitary-adrenal axis are thought to play a role in the aetiology of major depression.*) We sought to examine **this issue** in the elderly ...
mixed	18.2%	(S (PP Given these data) (, ,) (**NP** *decreasing HTD to* $<$ *or* $=$ 5 *years*) (**VP** *may have a detrimental effect on patients with locally advanced prostate cancer*) (. .)) Only a randomized trial will conclusively clarify **this issue**.
nominalization	17.4%	As (**NP** *the influence of estrogen alone on breast cancer detection*) is not established, we examined **this issue** in the Women's Health Initiative trial...

Table 1: Antecedent types. In examples, the antecedent type is in **bold** and the marked antecedent is in *italics*.

by the annotator h by r_{hj}. In Figure 1, annotators 1 and 2 have reached different conclusions by identifying 9 and 10 sections respectively in their copies of the text. Annotator 1 has not marked any antecedent for the anaphor with $id = 1$, while annotator 2 has marked r_{21} for the same anaphor. Both annotators have marked exactly the same antecedent for the anaphor with $id = 4$. The difference between two annotated sections is defined in terms of the square of the distance between the non-overlapping parts of the sections. The distance is 0 when the sections are unmarked by both annotators or are marked and exactly same, and is the summation of the squares of the unmatched parts if they are different. The coefficient is computed using intersections of the marked sections. In Figure 1, annotators 1 and 2 have a total of 14 intersections. The observed disagreement ${}_uD_o$ is the weighted sum of the differences between all mismatching intersections of sections marked by the annotators, and the expected disagreement is the summation of all possible differences of pairwise combinations of all sections of all annotators normalized by the length of the text (in terms of the number of tokens) and the number of pairwise combinations of annotators.

For our data, the inter-annotator agreement was $\alpha_u = 0.86$ (${}_uD_o = 0.81$ and ${}_uD_e = 5.81$) despite the fact that the annotators differed in their domain expertise, which suggests that abstract concepts such as *issue* can be annotated reliably.

3.2 Corpus Statistics

A gold standard corpus was created by resolving the cases where the annotators disagreed. Among 132 training instances, the annotators could not resolve 6 instances and we broke the tie by writing to the authors of the articles and using their response to resolve the disagreement. In the gold standard corpus, 95.5% of the antecedents were in the current or previous sentence and 99.2% were in the current or previous two sentences. Only one antecedent was found more than two sentences back and it was six sentences back. One instance was a cataphor, but the antecedent occurred in the same sentence as the anaphor. This suggests that for an automatic *this-issue* resolution system, it would be reasonable to consider only the previous two sentences along with the sentence containing the anaphor.

The distribution of the different linguistic forms that an antecedent of *this-issue* can take in our data set is shown in Table 1. The majority of antecedents are clauses or whole sentences. A number of antecedents are noun phrases, but these are generally nominalizations that refer to abstract concepts (e.g., *the influence of estrogen alone on breast cancer detection*). Some antecedents are not even well-defined syntactic constituents[5] but are combinations of several well-defined constituents. We denote the type of such antecedents as *mixed*. In the corpus, 18.2% of the antecedents are of this type, suggesting that it is not sufficient to restrict the antecedent search space to well-defined syntactic constituents.[6]

In our data, we did not find anaphoric chains for any of the *this-issue* anaphor instances, which indicates that the antecedents of *this-issue* anaphors are

[5]We refer to every syntactic constituent identified by the parser as a *well-defined syntactic constituent*.

[6]Indeed, many of mixed type antecedents (nearly three-quarters of them) are the result of parser attachment errors, but many are not.

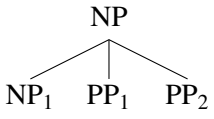
in the reader's local memory and not in the global memory. This observation supports the THIS-NPs hypothesis (Gundel et al., 1993; Poesio and Modjeska, 2002) that *this*-NPs are used to refer to entities which are *active* albeit not in *focus*, i.e., they are not the center of the previous utterance.

4 Resolution Algorithm

4.1 Candidate Extraction

For correct resolution, the set of extracted candidates must contain the correct antecedent in the first place. The problem of candidate extraction is non-trivial in abstract anaphora resolution because the antecedents are of many different types of syntactic constituents such as clauses, sentences, and nominalizations. Drawing on our observation that the mixed type antecedents are generally a combination of different well-defined syntactic constituents, we extract the set of candidate antecedents as follows. First, we create a set of candidate sentences which contains the sentence containing the *this-issue* anaphor and the two preceding sentences. Then, we parse every candidate sentence with the Stanford Parser[7]. Initially, the set of candidate constituents contains a list of well-defined syntactic constituents. We require that the node type of these constituents be in the set {S, SBAR, NP, SQ, SBARQ, S+V}. This set was empirically derived from our data. To each constituent, there is associated a set of mixed type constituents. These are created by concatenating the original constituent with its sister constituents. For example, in (4), the set of well-defined eligible candidate constituents is {NP, NP_1} and so NP_1 PP_1 is a mixed type candidate.

(4)



The set of candidate constituents is updated with the extracted mixed type constituents. Extracting mixed type candidate constituents not only deals with mixed type instances as shown in Table 1, but as a side effect it also corrects some attachment errors made by the parser. Finally, the constituents having a number of leaves (words) less than a threshold[8] are discarded to give the final set of candidate constituents.

4.2 Features

We explored the effect of including 43 automatically extracted features (12 feature classes), which are summarized in Table 2. The features can also be broadly divided into two groups: *issue*-specific features and general abstract-anaphora features. *Issue*-specific features are based on our common-sense knowledge of the concept of *issue* and the different semantic forms it can take; e.g., controversy (*X is controversial*), hypothesis (*It has been hypothesized X*), or lack of knowledge (*X is unknown*), where *X* is the issue. In our data, we observed certain syntactic patterns of issues such as *whether X or not* and *that X* and the IP feature class encodes this information. Other *issue*-specific features are IVERB and IHEAD. The feature IVERB checks whether the governing verb of the candidate is an issue verb (e.g., *speculate*, *hypothesize*, *argue*, *debate*), whereas IHEAD checks whether the candidate head in the dependency tree is an issue word (e.g., *controversy*, *uncertain*, *unknown*). The general abstract-anaphora resolution features do not make use of the semantic properties of the word *issue*. Some of these features are derived empirically from the training data (e.g., ST, L, D). The EL feature is borrowed from Müller (2008) and encodes the embedding level of the candidate within the candidate sentence. The MC feature tries to capture the idea of the THIS-NPs hypothesis (Gundel et al., 1993; Poesio and Modjeska, 2002) that the antecedents of *this*-NP anaphors are not the center of the previous utterance. The general abstract-anaphora features in the SR feature class capture the semantic role of the candidate in the candidate sentence. We used the Illinois Semantic Role Labeler[9] for SR features. The general abstract-anaphora features also contain a few lexical features (e.g., M, SC). But these features are independent of the semantic properties of the word *issue*. The general abstract-anaphora resolution features also contain dependency-tree features, lexical-

[7] http://nlp.stanford.edu/software/lex-parser.shtml

[8] The threshold 5 was empirically derived. Antecedents in our training data had on average 17 words.

[9] http://cogcomp.cs.illinois.edu/page/software_view/SRL

ISSUE PATTERN (IP)	
ISWHETHER	1 iff the candidate follows the pattern `SBAR` → `(IN whether) (S ...)`
ISTHAT	1 iff the candidate follows the pattern `SBAR` → `(IN that) (S ...)`
ISIF	1 iff the candidate follows the pattern `SBAR` → `(IN iff) (S ...)`
ISQUESTION	1 iff the candidate node is `SBARQ` or `SQ`
SYNTACTIC TYPE (ST)	
ISNP	1 iff the candidate node is of type NP
ISS	1 iff the candidate node is a sentence node
ISSBAR	1 iff the candidate node is an SBAR node
ISSQ	1 iff the candidate node is an SQ or SBARQ node
MIXED	1 iff the candidate node is of type *mixed*
EMBEDDING LEVEL (EL) (Müller, 2008)	
TLEMBEDDING	level of embedding of the given candidate in its top clause (the root node of the syntactic tree)
ILEMBEDDING	level of embedding of the given candidate in its immediate clause (the closest parent of type S or SBAR)
MAIN CLAUSE (MC)	
MCLAUSE	1 iff the candidate is in the main clause
DISTANCE (D)	
ISSAME	1 iff the candidate is in the same sentence as anaphor
SADJA	1 iff the candidate is in the adjacent sentence
ISREM	1 iff the candidate occurs 2 or more sentences before the anaphor
POSITION	1 iff the antecedent occurs before anaphor
SEMANTIC ROLE LABELLING (SR)	
IVERB	1 iff the governing verb of the given candidate is an *issue* verb
ISA0	1 iff the candidate is the *agent* of the governing verb
ISA1	1 iff the candidate is the *patient* of the governing verb
ISA2	1 iff the candidate is the *instrument* of the governing verb
ISAM	1 iff the candidate plays the role of *modiffication*
ISNOR	1 iff the candidate plays no well-defined semantic role in the sentence
DEPENDENCY TREE (DT)	
IHEAD	1 iff the candidate head in the dependency tree is an issue word (e.g., *controversial*, *unknown*)
ISSUBJ	1 iff the dependency relation of the candidate to its head is of type *nominal, controlling* or *clausal subject*
ISOBJ	1 iff the dependency relation of the candidate to its head is of type *direct object* or *preposition obj*
ISDEP	1 iff the dependency relation of the candidate to its head is of type *dependent*
ISROOT	1 iff the candidate is the root of the dependency tree
ISPREP	1 iff the dependency relation of the candidate to its head is of type *preposition*
ISCONT	1 iff the dependency relation of the candidate to its head is of type *continuation*
ISCOMP	1 iff the dependency relation of the candidate to its head is of type *clausal* or *adjectival complement*
ISSENT	1 iff candidate's head is the root node
PRESENCE OF MODALS (M)	
MODAL	1 iff the given candidate contains a modal verb
PRESENCE OF SUBORDINATING CONJUNCTION (SC)	
ISCONT	1 iff the candidate starts with a contrastive subordinating conjunction (e.g., *however*, *but*, *yet*)
ISCAUSE	1 iff the candidate starts with a causal subordinating conjunction (e.g., *because*, *as*, *since*)
ISCOND	1 iff the candidate starts with a conditional subordinating conjunction (e.g., *if*, *that*, *whether or not*)
LEXICAL OVERLAP (LO)	
TOS	normalized ratio of the overlapping words in candidate and the title of the article
AOS	normalized ratio of the overlapping words in candidate and the anaphor sentence
DWS	proportion of domain-specific words in the candidate
CONTEXT (C)	
ISPPREP	1 iff the preceding word of the candidate is a preposition
ISFPREP	1 iff the following word of the candidate is a preposition
ISPPUNCT	1 iff the preceding word of the candidate is a punctuation
ISFPUNCT	1 iff the following word of the candidate is a punctuation
LENGTH (L)	
LEN	length of the candidate in words

Table 2: Feature sets for *this-issue* resolution. All features are extracted automatically.

overlap features, and context features.

4.3 Candidate Ranking Model

Given an anaphor a_i and a set of candidate antecedents $C = \{C_1, C_2, ..., C_k\}$, the problem of anaphora resolution is to choose the best candidate antecedent for a_i. We follow the candidate-ranking model proposed by Denis and Baldridge (2008). The advantage of the candidate-ranking model over the mention-pair model is that it overcomes the strong independence assumption made in mention-pair models and evaluates how good a candidate is relative to *all* other candidates.

We train our model as follows. If the anaphor is a *this-issue* anaphor, the set C is extracted using the candidate extraction algorithm from Section 4.1. Then a corresponding set of feature vectors, $C_f = \{C_{f1}, C_{f2}, ..., C_{fk}\}$, is created using the features in Table 2. The training instances are created as described by Soon et al. (2001). Note that the instance creation is simpler than for general coreference resolution because of the absence of anaphoric chains in our data. For every anaphor a_i and eligible candidates $C_f = \{C_{f1}, C_{f2}, ..., C_{fk}\}$, we create training examples $(a_i, C_{fi}, label), \forall C_{fi} \in C_f$. The label is 1 if C_i is the true antecedent of the anaphor a_i, otherwise the label is -1. The examples with label 1 get the rank of 1, while other examples get the rank of 2. We use SVM^{rank} (Joachims, 2002) for training the candidate-ranking model. During testing, the trained model is used to rank the candidates of each test instance of *this-issue* anaphor.

5 Evaluation

In this section we present the evaluation of each component of our resolution system.

5.1 Evaluation of Candidate Extraction

The set of candidate antecedents extracted by the method from Section 4.1 contained the correct antecedent 92% of the time. Each anaphor had, on average, 23.80 candidates, of which only 5.19 candidates were nominal type. The accuracy dropped to 84% when we did not extract mixed type candidates. The error analysis of the 8% of the instances where we failed to extract the correct antecedent revealed that most of these errors were parsing errors which could not be corrected by our candidate extraction method.[10] In these cases, the parts of the antecedent had been placed in completely different branches of the parse tree. For example, in (5), the correct antecedent is a combination of the NP from the $S \rightarrow VP \rightarrow NP \rightarrow PP \rightarrow$ **NP** branch and the PP from $S \rightarrow VP \rightarrow$ **PP** branch. In such a case, concatenating sister constituents does not help.

(5) The data from this pilot study (VP (VBP provide) (NP (NP no evidence) (PP (IN for) (NP **a difference in hemodynamic effects between pulse HVHF and CPFA**))) (PP **in patients with septic shock already receiving CRRT**)). A larger sample size is needed to adequately explore **this issue**.

5.2 Evaluation of *this-issue* Resolution

We propose two metrics for abstract anaphora evaluation. The simplest metric is the percentage of antecedents on which the system and the annotated gold data agree. We denote this metric as *EXACT-M* (Exact Match) and compute it as the ratio of number of correctly identified antecedents to the total number of marked antecedents. This metric is a good indicator of a system's performance; however, it is a rather strict evaluation because, as we noted in section 1, issues generally have no precise boundaries in the text. So we propose another metric called RLL, which is similar to the ROUGE-L metric (Lin, 2004) used for the evaluation of automatic summarization. Let the marked antecedents of the gold corpus for k anaphor instances be $G = \langle g_1, g_2, ..., g_k \rangle$ and the system-annotated antecedents be $A = \langle a_1, a_2, ..., a_k \rangle$. Let the number of words in G and A be m and n respectively. Let $LCS(g_i, a_i)$ be the the number of words in the longest common subsequence of g_i and a_i. Then the precision (P_{RLL}) and recall (R_{RLL}) over the whole data set are computed as shown in equations (2) and (3). P_{RLL} is the total number of word overlaps between the gold and system-annotated antecedents normalized by the number of words in system-annotated antecedents and R_{RLL} is the total number of such word overlaps normalized by the number of words in the gold antecedents. If the system picks too much text for antecedents, R_{RLL} is high but P_{RLL} is low. The F-score,

[10] Extracting candidate constituents from the dependency trees did not add any new candidates to the set of candidates.

		5-fold Cross-Validation				Test			
		P_{RLL}	R_{RLL}	F_{RLL}	EX-M	P_{RLL}	R_{RLL}	F_{RLL}	EX-M
1	Adjacent sentence	66.47	86.16	74.93	22.93	61.73	87.69	72.46	24.00
2	Random	50.71	32.84	39.63	8.40	43.75	35.00	38.89	15.69
3	{IP, D, C, LO, EL, M, MC, L, SC, SR, DT}	79.37	83.66	81.11	59.80	71.89	85.74	78.20	58.82
4	{IP, D, C, LO, M, MC, L, SC, DT}	78.71	83.86	81.14	59.89	70.64	88.09	78.40	54.90
5	{IP, D, C, EL, L, SC, SR, DT}	77.95	83.06	80.33	57.41	72.03	84.85	77.92	**60.78**
6	{IP, D, EL, MC, L, SR, DT}	80.00	84.75	**82.24**	**59.91**	68.88	85.29	76.22	56.86
7	{IP, D, M, L, SR}	73.42	83.16	77.90	52.31	70.74	91.03	**79.61**	50.98
8	{D, C, LO, L, SC, SR, DT}	79.15	85.28	82.04	56.07	67.39	86.32	75.69	52.94
9	*issue*-specific features	74.66	45.70	56.57	41.42	64.20	45.88	53.52	41.38
10	non-*issue* features	76.39	79.39	77.82	51.48	71.19	83.24	76.75	58.82
11	All	78.22	82.92	80.41	56.75	71.28	83.24	76.80	56.86
12	Oracle candidate extractor + row 3	79.63	82.26	80.70	58.32	74.65	87.06	80.38	**64.71**
13	Oracle candidate sentence extractor + row 3	86.67	92.12	**89.25**	**63.72**	79.71	91.49	**85.20**	62.00

Table 3: *this-issue* resolution results with SVM^{rank}. *All* means evaluation using all features. *Issue*-specific features = {IP, IVERB, IHEAD}. EX-M is EXACT-M.

F_{RLL}, combines these two scores.

$$P_{RLL} = \frac{1}{n}\sum_{i=1}^{k} LCS(g_i, a_i) \quad (2)$$

$$R_{RLL} = \frac{1}{m}\sum_{i=1}^{k} LCS(g_i, a_i) \quad (3)$$

$$F_{RLL} = \frac{2 \times P_{RLL} \times R_{RLL}}{P_{RLL} + R_{RLL}} \quad (4)$$

The lower bound of F_{RLL} is 0, where no true antecedent has any common substring with the predicted antecedents and the upper bound is 1, where all the predicted and true antecedents are exactly the same. In our results we represent these scores in terms of percentage.

There are no implemented systems that resolve *issue* anaphora or abstract anaphora signalled by label nouns in arbitrary text to use as a comparison. So we compare our results against two baselines: *adjacent sentence* and *random*. The adjacent sentence baseline chooses the previous sentence as the correct antecedent. This is a high baseline because in our data 84.1% of the antecedents lie within the adjacent sentence. The random baseline chooses a candidate drawn from a uniform random distribution over the set of candidates.[11]

[11] Note that our F_{RLL} scores for both baselines are rather high because candidates often have considerable overlap with one another; hence a wrong choice may still have a high F_{RLL} score.

We carried out two sets of systematic experiments in which we considered all combinations of our twelve feature classes. The first set consists of 5-fold cross-validation experiments on our training data. The second set evaluates how well the model built on the training data works on the unseen test data.

Table 3 gives results of our system. The first two rows are the baseline results. Rows 3 to 8 give results for some of the best performing feature sets. All systems based on our features beat both baselines on F-scores and EXACT-M. The empirically derived feature sets IP (issue patterns) and D (distance) appeared in almost all best feature set combinations. Removing D resulted in a 6 percentage points drop in F_{RLL} and a 4 percentage points drop in EXACT-M scores. Surprisingly, feature set ST (syntactic type) was not included in most of the best performing set of feature sets. The combination of syntactic and semantic feature sets {IP, D, EL, MC, L, SR, DT} gave the best F_{RLL} and EXACT-M scores for the cross-validation experiments. For the test-data experiments, the combination of semantic and lexical features {D, C, LO, L, SC, SR, DT} gave the best F_{RLL} results, whereas syntactic, discourse, and semantic features {IP, D, C, EL, L, SC, SR, DT} gave the best EXACT-M results. Overall, row 3 of the table gives reasonable results for both cross-validation and test-data experiments with no statistically significant difference to the corresponding best

EXACT-M scores in rows 6 and 5 respectively.[12]

To pinpoint the errors made by our system, we carried out three experiments. In the first experiment, we examined the contribution of *issue*-specific features versus non-*issue* features (rows 9 and 10). Interestingly, when we used only non-*issue* features, the performance dropped only slightly. The F_{RLL} results from using only *issue*-specific features were below baseline, suggesting that the more general features associated with abstract anaphora play a crucial role in resolving *this-issue* anaphora.

In the second experiment, we determined the error caused by the candidate extractor component of our system. Row 12 of the table gives the result when an oracle candidate extractor was used to add the correct antecedent in the set of candidates whenever our candidate extractor failed. This did not affect cross-validation results by much because of the rarity of such instances. However, in the test-data experiment, the EXACT-M improvements that resulted were statistically significant. This shows that our resolution algorithm was able to identify antecedents that were arbitrary spans of text.

In the last experiment, we examined the effect of the reduction of the candidate search space. We assumed an oracle candidate sentence extractor (Row 13) which knows the exact candidate sentence in which the antecedent lies. We can see that both RLL and EXACT-M scores markedly improved in this setting. In response to these results, we trained a decision-tree classifier to identify the correct antecedent sentence with simple location and length features and achieved 95% accuracy in identifying the correct candidate sentence.

6 Discussion and Conclusions

We have demonstrated the possibility of resolving complex abstract anaphora, namely, *this-issue* anaphora having arbitrary antecedents. The work takes the annotation work of Botley (2006) and Dipper and Zinsmeister (2011) to the next level by resolving *this-issue* anaphora automatically. We proposed a set of 43 automatically extracted features that can be used for resolving abstract anaphora.

Our results show that general abstract-anaphora resolution features (i.e., other than *issue*-specific features) play a crucial role in resolving *this-issue* anaphora. This is encouraging, as it suggests that the approach could be generalized for other NPs — especially NPs having similar semantic constraints such as *this problem*, *this decision*, and *this conflict*.

The results also show that reduction of search space markedly improves the resolution performance, suggesting that a two-stage process that first identifies the broad region of the antecedent and then pinpoints the exact antecedent might work better than the current single-stage approach. The rationale behind this two-stage process is twofold. First, the search space of abstract anaphora is large and noisy compared to nominal anaphora.[13] And second, it is possible to reduce the search space and accurately identify the broad region of the antecedents using simple features such as the location of the anaphor in the anaphor sentence (e.g., if the anaphor occurs at the beginning of the sentence, the antecedent is most likely present in the previous sentence).

We chose scientific articles over general text because in the former domain the actual referents are seldom discourse deictic (i.e., not present in the text). In the news domain, for instance, which we have also examined and are presently annotating, a large percentage of *this-issue* antecedents lie outside the text. For example, newspaper articles often quote sentences of others who talk about the issues in their own world, as shown in example (6).

(6) As surprising and encouraging to organizers of the movement are the Wall Street names added to their roster. Prominent among them is Paul Singer, a hedge fund manager who is straight and chairman of the conservative Manhattan Institute. He has donated more than $8 million to various same-sex marriage efforts, in states including California, Maine, New Hampshire, New Jersey, New York and Oregon, much of it since 2007.

"It's become something that gradually peo-

[12] We performed a simple one-tailed, k-fold cross-validated paired t-test at significance level $p = 0.05$ to determine whether the difference between the EXACT-M scores of two feature classes is statistically significant.

[13] If we consider all well-defined syntactic constituents of a sentence as issue candidates, in our data, a sentence has on average 43.61 candidates. Combinations of several well-defined syntactic constituents only add to this number. Hence if we consider the antecedent candidates from the previous 2 or 3 sentences, the search space can become quite large and noisy.

ple like myself weren't afraid to fund, weren't afraid to speak out on," Mr. Singer said in an interview. "I'm somebody who is philosophically very conservative, and on **this issue** I thought that this really was important on the basis of liberty and actual family stability."

In such a case, the antecedent of *this issue* is not always in the text of the newspaper article itself, but must be inferred from the context of the quotation and the world of the speaker quoted. That said, we do not use any domain-specific information in our *this-issue* resolution model. Our features are solely based on distance, syntactic structure, and semantic and lexical properties of the candidate antecedents which could be extracted for text in any domain.

Issue anaphora can also be signalled by demonstratives other than *this*. However, for our initial study, we chose *this issue* for two reasons. First, in our corpus as well as in other general corpora such as the New York Times corpus, *issue* occurs much more frequently with *this* than other demonstratives. Second, we did not want to increase the complexity of the problem by including the plural *issues*.

Our approach needs further development to make it useful. Our broad goal is to resolve abstract anaphora signalled by label nouns in all kinds of text. At present, the major obstacle is that there is very little annotated data available that could be used to train an abstract anaphora resolution system. And the understanding of abstract anaphora itself is still at an early stage; it would be premature to think about unsupervised approaches. In this work, we studied the narrow problem of resolution of *this-issue* anaphora in the medical domain to get a good grasp of the general abstract-anaphora resolution problem.

A number of extensions are planned for this work. First, we will extend the work to resolve other abstract anaphors (e.g., *this decision*, *this problem*). Second, we will experiment with a two-stage resolution approach. Third, we would like to explore the effect of including serious discourse structure features in our model. (The feature sets SC and C encode only shallow discourse information.) Finally, during annotation, we noted a number of *issue* patterns (e.g., *An open question is X*, *X is under debate*); a possible extension is extracting issues and problems from text using these patterns as seed patterns.

7 Acknowledgements

We thank Dr. Brian Budgell from the Canadian Memorial Chiropractic College for annotating our data and for helpful discussions. We also thank the anonymous reviewers for their detailed and constructive comments. This research was financially supported by the Natural Sciences and Engineering Research Council of Canada and by the University of Toronto.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Philip Simon Botley. 2006. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.

Donna K. Byron. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. *Technical Report, University of Rochester*.

Donna K. Byron. 2004. *Resolving pronominal reference to abstract entities*. Ph.D. thesis, Rochester, New York: University of Rochester.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for NLP*, Alicante, Spain, June.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. 2011. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, November.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Honolulu, Hawaii, October. Association for Computational Linguistics.

Stefanie Dipper and Heike Zinsmeister. 2011. Annotating abstract anaphora. *Language Resources and Evaluation*, 69:1–16.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17:51–89.

Gill Francis. 1994. Labelling discourse: an aspect of nominal group lexical cohesion. In Malcolm Coulthard, editor, *Advances in written text analysis*, pages 83–101, London. Routledge.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring ex-

pressions in discourse. *Language*, 69(2):274–307, June.

Graeme Hirst. 1981. *Anaphora in Natural Language Understanding: A Survey*, volume 119 of *Lecture Notes in Computer Science*. Springer.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142.

Klaus Krippendorff. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, 25:47–76.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition.

Yu-Hsiang Lin and Tyne Liang. 2004. Pronominal and sortal anaphora resolution for biomedical literature. In *Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*, Taiwan, September.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman.

Natalia N. Modjeska. 2003. *Resolving Other-Anaphora*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Christoph Müller. 2008. *Fully Automatic Resolution of It, This and That in Unrestricted Multi-Party Dialog*. Ph.D. thesis, Universität Tübingen.

Costanza Navarretta. 2011. Antecedent and referent types of abstract pronominal anaphora. In *Proceedings of the Workshop Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena*, Göttingen, Germany, February.

Rebecca Passonneau. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 51–59, Vancouver, British Columbia, Canada, June. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Massimo Poesio and Natalia N. Modjeska. 2002. The THIS-NPs hypothesis: A corpus-based investigation. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Conference (DAARC 2002)*, pages 157–162, Lisbon, Portugal, September.

Massimo Poesio, Simone Ponzetto, and Yannick Versley. 2011. Computational models of anaphora resolution: A survey. Unpublished.

Hans-Jörg Schmid. 2000. *English Abstract Nouns As Conceptual Shells: From Corpus to Cognition*. Topics in English Linguistics. De Gruyter Mouton, Berlin.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan, July. Association for Computational Linguistics.

Manabu Torii and K. Vijay-Shanker. 2007. Sortal anaphora resolution in Medline abstracts. *Computational Intelligence*, 23(1):15–27.