Anaphora in a Wider Context: Tracking Discourse Referents

Christopher Kennedy[1] **and Branimir Boguraev**[2]

Abstract. A number of linguistic and stylistic devices are employed in text-based discourse for the purposes of introducing, defining, refining, and re-introducing discourse entities. This paper looks at one of the most pervasive of these mechanisms, anaphora, and addresses the question of how current computational approaches to anaphora scale up to building, and maintaining, a richer model of text structure, which embodies the notion of a discourse referent's behaviour in the entire text. Given the less than fully robust status of syntactic parsers to date, we question the applicability of current anaphora resolution algorithms to open-ended text types, styles, and genres. We outline an algorithm for anaphora resolution, which modifieds and extends a configurationally-based approach, while working from the output of a part of speech tagger, enriched only with annotations of grammatical function. Without compromising output quality, the algorithm compensates for the shallower level of analysis with mechanisms for identifying different text forms for each discourse referent, and for maintaining awareness of inter-sentential context. A salience measure—for each discourse referent, over the entire text—not only crucially drives the algorithm, but also effectively maintains a record of where and how discourse referents occur in the text. Anaphora resolution thus becomes an integral part of a deeper discourse analysis process, ultimately concerned with tracking discourse referents.

1 ANAPHORA IN A WIDER CONTEXT

A core question in computational discourse modelling concerns the identification and representation of discourse referents: the actors and objects around which a story unfolds. In general, there are two sides to this: identifying the ways in which the same entity can be referred to, and establishing that a set of potentially coreferential 'text objects' which are in fact so.

A number of linguistic devices come to play when a reference to a previously introduced object needs to be established, and the complexity and range of such devices is considerable. For the purposes of practical natural language processing, not all of these have been given equal attention. For instance, work on text analysis and content extraction has tended to focus extensively on naming and abbreviatory conventions (e.g., the conditions under which "American National Standards Institute", "the institute", and "ANSI" could be co-referential in a document); more detailed discussion of such topics can be found in [7] and [8].

In fact, a whole class of text processing applications—aiming to account for a particular style of news reporting—have recently addressed the question of discourse referent co-referentiality, with a strong emphasis on normalising variance in referring to actors in the discourse: the example below (due to S. Nirenburg), illustrates some of the complexities involved in establishing coreferentiality among the emphasized phrases.

> PRIEST IS CHARGED WITH POPE ATTTACK
>
> *A Spanish Priest* was charged here today with attempting to murder the Pope. *Juan Fernandez Krohn*, aged 32, was arrested after *a man armed with a bayonet* approached the Pope while he was saying prayers at Fatima on Wednesday night.
>
> According to the police, *Fernandez* told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope 'looked furious' on hearing *the priest's* criticism of his handling of the church's affairs. If found guilty, *the Spaniard* faces a prison sentence of 15–20 years.

Another aspect of the problem has become prominent in recent work on terminology identification. The argument, first put forward in [3], that technical terms are defined as noun phrases with certain discourse properties, gives rise to algorithms for extracting scientific terminology, as well as for general indexing purposes. For optimal performance, it is clearly essential that text analysis procedures be capable of 'normalizing' reduced forms of terms to their correct canonical 'base': consider, for instance, a technical manual in the domain of hard storage maintenance, where a mention of "the disk" could equally well refer to "floppy disk", "internal hard disk", or "RAM disk", and is only interpretable in context. More detailed discussion of these issues, and a particular interpretation strategy, can be found in [2].

Most pervasive, however, and common to all types of text and genre, is the phenomenon of anaphoric reference. Usually tackled in the context of a machine translation task, the fact remains that no strong procedure for discourse model building can be devised without a robust anaphora resolution component. Work on computational anaphora resolution to date has tended to assume full syntactic analysis of the text as a base: thus only a relatively small class of text processing applications would have access to sophisticated mechanisms for resolving pronominal references.

Our concern is with the general problem of delivery of content analysis to a depth involving non-trivial amount of discourse processing including, but not limited to, anaphora resolution. We disallow assumptions concerning domain, style, and genre of input—in effect, imposing a requirement not to rely exclusively on full (configurational) syntactic analysis. To this end, we have been working on a text processing framework which builds its capabilities entirely on the basis of a shallow (non-configurational) linguistic analysis of the input stream, thus trading off depth of base level analysis for breadth of coverage. The question of overall strategy for supplying the higher-level semantic and pragmatic processes with sufficient linguistic information has been discussed, to some extent, in [2]: in

[1] Board of Studies in Linguistics, University of California, Santa Cruz, CA 95064, kennedy@ling.ucsc.edu

[2] Advanced Technologies Group, Apple Computer, Inc., Cupertino, CA 95014, bkb@apple.com

ECAI 96. *12th European Conference on Artificial Intelligence*
Edited by W. Wahlster
Published in 1996 by John Wiley & Sons, Ltd.

summary, the argument is that such a strategy should be grounded in an exploitation of lexically-intensive analysis of arbitrary text to implement what is, in essence, a strongly semantic task.

The focus of this paper is on anaphora resolution as an essential prerequisite to building a discourse model of text. In the light of the preceding remarks, our attention is focused on two areas. First, we address the problem of working from a shallower linguistic base. For the underlying capability of pronominal anaphora resolution, we build upon an algorithm with high rate of correct analysis, presented in [6]. While one of its strongest points is that it operates primarily on syntactic information alone, this is also a limiting factor for its wide use: current state-of-the-art of parsing technology still falls short of delivery, robustly and reliably, of syntactic analysis of real texts to the level of detail and precision required by the filters and constraints of Lappin and Leass.

Next, we look at anaphoric reference specifically as a device for following the discourse salience of reference objects, and observe that for this, anaphora resolution must be sensitive to context larger than the hitherto postulated window of not more than several sentences. In principle, the resolution algorithm ought to be able to identify references to the same entity even if these are separated by the entire span of a document. While it is unrealistic to assume that a simple pronominal mention would be directly resolvable to a referent introduced pages earlier, it is certainly the case that, by identifying correctly its (recent) referent, we could—and should—establish coreferentiality with that same referent as it is brought into prominence in all of its mentions in the text. If analyzing anaphors is done as part of building an extended discourse model, then the analysis process needs to be aware of the entire document span.

Below, we describe the modified Lappin/Leass algorithm in some detail. We assume some acquaintance with the original version, presented in [6]; see also [5] for more detaled discussion of our implementation. For the purposes of this paper, we particularly focus on demonstrating how the characteristics of the shallow linguistic analysis necessitate adjusting some of the parameters of the algorithm, as well as how the set of filters and constraints of the original algorithm need to be re-cast in order to account for the new form of input.

We also discuss some additions to the input analysis, in particular a much stronger awareness of inter-sentential context, which enrich the informational base of the anaphora resolution process. As we will argue, this is not just an enhancement to the original algorithm, which happens to contribute to the overall accuracy of our output. Rather, it is a necessary adjustment, in the light of the requirement for extending anaphora to a wider context. We elaborate the notion of a "discourse referent", as a generalized representation for discourse entities distributed in the text, and demonstrate how continued awareness of the discourse properties of *each* discourse referent translates into an overall measure of salience; this, in its own turn, allows us to track discourse entities as the story unfolds.

2 DISCOURSE REFERENTS

The base level linguistic analysis for anaphora resolution is the output of a part-of-speech tagger, augmented with syntactic function annotations for each input token: this kind of analysis is generated by the morphosyntactic tagging system described in [4], [9] (henceforth LINGSOFT). In addition to extremely high levels of accuracy in recall and precision of tag assignment ([9] reports 99.77% overall recall and 95.54% overall precision), the primary motivation for adopting this system is the requirement to develop a robust text processor—with anaphora resolution being just one of its discourse analysis functions—capable of handling arbitrary real input reliably.

The modified algorithm we present requires additional annotation of the input text stream by a simple position-identification function which assigns to each text token an integer value representing its offset in the stream. The tagger provides a very simple analysis of the structure of the text, annotating lexical items with morphological, lexical, grammatical and syntactic features. As an example, given the text (fragment from a press release announcement)

> "IISP, which consists of standards developing organizations, industry associations, consortia and architecture groups, companies and government, held its first meeting in New York last July."

the input for the anaphora resolution algorithm would be:

```
"IISP/off215" "IISP" N NOM SG @SUBJ
"$\,/off216"
"which/off217" "which" PRON WH NOM SG/PL @SUBJ
"consists/off218" "consist" V PRES SG3 VFIN @+FMAINV
"of/off219" "of" PREP  @ADVL
"standards/off220" "standard" N NOM PL @<P
"developing/off221" "develop" PCP1 @<NOM-FMAINV
"organizations/off222" "organization" N NOM PL @OBJ
"$\,/off223"
........
"companies/off232" "company" N NOM PL @SUBJ @OBJ @<P
"and/off233" "and" CC @CC
"government/off234" "government" N NOM SG/PL @SUBJ @OBJ @<P
"$\,/off235"
"held/off236" "hold" V PAST VFIN @+FMAINV
"its/off237" "it" PRON GEN SG3 @GN>
"first/off238" "first" NUM ORD @DN>
"meeting/off239" "meet" PCP1  @OBJ @-FMAINV
"in/off240" "in" PREP @<NOM @ADVL
"NewYork/off241" "New_York"  N NOM SG @<P
........
```

Each lexical token has syntactic function information associated with it: thus "IISP" has been analyzed as a subject (@SUBJ), and "held" has been identified as a main verb. Although LINGSOFT does not provide specific information about constituent structure, partial constituency—specifically, identification of sequences of tokens as phrasal units such as, for instance, modifier-head-complement sequences—can be inferred by applying a set of filters to the tagged text. These are defined as patterns, stated as regular expressions over metatokens inthe tagger output. Note that in addition to being able to derive phrasal level analysis, the information in the output stream makes it possible to infer some labelling information for constituents (e.g. stamping a noun phrase with its grammatical function, and/or indicating whether it was observed in an adjunct).

2.1 Mining for discourse referents

The primary data for anaphora resolution is a complete set of all discourse referents. As these are realized as noun phrases, the set is derived by a phrasal grammar, whose patterns characterize the composition of an NP in terms of possible token sequences.

The NP identification patterns are divided into two sets. The first, more general, set identifies token sequences corresponding to noun phrases. The second set of patterns detects nominal sequences in specific syntactic environments, in particular, in contexts involving containment in other nominal expressions or containment in adverbial adjuncts. The syntactic pattern matching annotates the NP phrases with the type of syntactic context they appear in (e.g. containment in an adverbial adjunct, in a prepositional complement of a noun, or in a clausal complement of noun). As we discuss in 2.2, the inevitable incorrect analyses of NP PP sequences due to the limited expressiveness of regular expressions, cand be filtered by later heuristics other than those which rely on syntactic context.

After unifying the (local) syntactic information with the (global) contextual cues, on a per-phrase basis, we arrive at a set of *discourse referents*—abstract objects behind the linguistic expressions referring to the participants in events described in the text. The following data is associated with each discourse referent: textual occurrence, specification of the type of the expression (referential or anaphoric), specification of the morphological features of the head of the phrase (including, for instance, data on agreement), the grammatical function of the constituent (as determined by LINGSOFT), information concerning the syntactic environment of the referential phrase, and positional information.

Starting the resolution process with a complete set of discourse referents is a direct consequence of the need to couch the resolution results not in terms of simple co-reference (manifested typically in pairing of an anaphor with its direct antecedent), but in terms of *co-reference classes*, where a class is defined as the set of all discourse referents sharing the same prototype. The Lappin/Leass algorithm, in fact, is easily adaptable to this notion of co-reference.

One of the required modifications of the algorithm needs to take into account co-referential effects over the entire text. To implement what is, in effect, a dynamically changing 'window of attention', within which local anaphoric filters operate most strongly (see *salience* in section 2.2 below), we augment the positional information for discourse referents to account for discourse segmentation effects. To this end, prior to anaphora resolution proper, the text is segmented into a set of local *contexts*. A context is defined to be a text fragment with high degree of coherence; context boundaries fall at perceptible breaks in the continuity of coherence. A number of strategies exist for performing such segmentation; our algorithm makes use of the procedure originally described in [1].

Note that each discourse referent contains information about itself *and* the context in which it appears, but the only information about its relation to other discourse referents is information about linear relations: precedence relations are determinable from offset value, but command relations are not explicitly encoded. This is an important point—and one of the essential differences from the Lappin/Leass algorithm. Lappin and Leass rely crucially on a set of disjoint reference filters, stated in terms of information about configurational relations between the discourse referents in a sentence. The modifications required for our implementation seek to provide a similar account, but inferred from a different, less rich, base.

2.2 Coreference and salience

Overall, the logic of the anaphora resolution algorithm parallels that of Lappin/Leass. Interpretation involves stepping through the text, a sentence at a time, and interpreting the discourse referents in the current sentence from left to right. A discourse referent either introduces a new concept or object into the discourse, or corefers with something else (already introduced earlier in the discourse). *Coreference* is determined by first locating those entities which an anaphoric expression cannot refer to, eliminating them from consideration, then finding the optimal antecedent among the remaining candidates; optimality is determined relative to a *salience* measure.

Coreference As in the Lappin/Leass algorithm, the general notion of coreference is represented in terms of equivalence classes of anaphorically related discourse referents, which we will refer to as *COREF classes*. A COREF class is defined as a set of discourse referents between which the algorithm has established a sequence of anaphoric links. The first discourse referent to mention a new concept triggers the introduction of a new COREF class; subsequent reference to the class by other discourse referents (as determined by the algorithm) causes new members to be added to the class. COREF classes are implemented as abstract objects which contain information about the set as a whole, including canonical form (typically determined by the discourse referent which introduces it), membership, and, most importantly: salience.

Salience The salience of a COREF class is determined by the status of its members with respect to 10 contextual, grammatical, and syntactic constraints. Following [6], we will refer to these as *salience factors*. Individual salience factors are associated with numerical values; the overall salience, or *salience weight* of a COREF is the sum of the values of the salience factors that are satisfied by some member of the COREF class (note that values may be satisfied at most once by each member of the class). Our salience factors are defined below with their values; they mirror those used by [6], with the exception of POSS-EMPHASIS (see below), and CONTEXT-RECENCY, which is sensitive to the *context*—i.e. the topically coherent segment of text—in which a discourse referent appears. Contexts are determined by a text-segmentation algorithm which follows [1]; this is the mechanism which accounts for the distributional pattern of the discourse referent, as it occurs through the entire text.

CONTEXT-RECENCY: (50) iff argument is in the current context.
SENTENCE-RECENCY: (100) iff argument is in the current sentence.
SUBJECT-EMPHASIS: (80) iff argument is a subject.
EXISTENTIAL-EMPHASIS: (70) iff argument is the pivot of an existential construction.
POSS-EMPHASIS: (65) iff argument is a possessive.
ACCUSATIVE-EMPHASIS: (50) iff argument is a direct object.
DATIVE-EMPHASIS: (40) iff argument is an indirect object.
OBLIQUE-EMPHASIS: (30) iff argument is contained in a PP.
HEAD-EMPHASIS: (80) iff argument is not contained in another NP.
ARGUMENT-EMPHASIS: (50) iff argument not contained in adjunct.

Note that the values of salience factors are arbitrary; what is crucial, as pointed out by [6], is the relational structure imposed on the factors by these values. The factors' relative ranking is justified both linguistically and by experimental results; [5] gives a more detailed account.

An important feature of our implementation of salience, following that of Lappin and Leass, is that it is variable: the salience of a COREF class decreases and increases according to the frequency of reference to the class. When an anaphoric link is established between a pronoun and a previously introduced discourse referent, the pronoun is added to the COREF class associated with the discourse referent, its COREF value is set to the COREF value of the antecedent (i.e., to the COREF object which represents the class), and the salience of the COREF object is recalculated according to how the new member satisfies the set of salience factors. This final step raises the overall salience of the COREF, since the new member will minimally satisfy SENTENCE-RECENCY and CONTEXT-RECENCY. In general, salience weight decreases, so that if new members are not added, the salience weight eventually reaches zero. Building in variable salience permits a realistic implementation of the *local* prominence of various COREF classes, allowing for a general heuristic for pronominal anaphora interpretation which resolves a pronoun to the most salient candidate antecedent. This procedure defines *local salience*.

As part of building a larger representation of text structure, however, it is essential that a measure is maintained of the saliency of a COREF class in relation to its prominence both in the text as a whole, and in the individual context(s) in which it accurs. Thus, a variation

on the computation above interprets the same conditions with respect to a non-decreasing salience weight. This does not replace the local prominence measure; rather, it reflects the distributional properties of a COREF class (and its members) as the text story unfolds, and provides an even more accurate account of tracking a particular discourse referent. The non-decreasing salience measure—*discourse salience*—enables the development of even more detailed models of discourse structure, overlayed onto the base segmentation strategy.

As we discussed in the introductory section, in order to maintain an accurate account of COREF class membership, it is necessary to be able to identify, reliably, different textual references to the same underlying discourse entity. The procedure described here is general enough to take into account, in addition to pronouns, other forms including contractions, abbreviations, and certain definite nominal compounds. The linguistic operations required for such interpretation are discussed in [2], and initial experimental results indicate that they may be directly integrated into the anaphora resolution process.

2.3 Anaphora resolution

Overall, the resolution procedure proper follows that of Lappin and Leass. The modifications to the original algorithm are fully described in [5]; this section only highlights some primary differences. In essence, there are three stages to interpreting discourse referents in a new sentence: adjustment of salience weights for COREFs, both existing and newly introduced; interpretetation of lexical anaphors (reflexives and reciprocals); and interpretation of pronouns.

Adjustment of salience weights is sensitive to the core distinction between COREFs being introduced in the current sentence and those already primed in the discourse. Lexical anaphors are resolved on the basis of a heuristic stipulating that a lexical anaphor must refer to a coargument. In the absence of configurational information, coarguments are identified using grammatical function information (as identified by the tagger). For instance, a lexical anaphor tagged as a DIRECT OBJECT is paired with the closest preceding discourse referent tagged as SUBJECT; similarly, possible antecedents for an INDIRECT OBJECT or OBLIQUE are derived from the closest preceding SUBJECT, as well as from a following OBJECT (as long as no other, embedded, SUBJECT intervenes). Multiple antecedents are ranked according to salience. Note that this operation makes heavy use of text precedence relations, in addition to syntactic function information.

The interpretation of pronouns proper follows the same basic resolution heuristic. However, extra care is required, as the generation of the candidate antecedents set needs to be sensitive to those discourse referents with which a pronoun cannot corefer. Disjoint reference is determined primarily by configurational relations between a pronoun and other constituents in the same sentence. In particular, a pronoun cannot corefer with a coargument, nor with a nonpronominal constituent which it commands and precedes, nor with a constituent which contains it. Without configurational syntactic information, we determine disjoint reference on the basis of inferences from grammatical function and precedence. These inferences are realized as a set of syntactic filters designed to determine co-argumentation, command, and containment. It turns out that even without specific information about constituent structure, the syntactic filters for disjoint reference are extremely accurate (see [5] for detailed discussion).

After the syntactic filters have been applied, a morphological filter further refines the list of possible antecedents, discarding discourse referents which disagree with the pronoun being interpreted. Additional structuring of the antecedents set is carried out by some methods for complex nominal interpretation, which follow [2]; these allow further 'collapsing' of apparently disjoint antecedent classes.

The discourse referents that remain after syntactic and morphological filtering form the set of candidate antecedents for the pronoun. A final set of evaluations is applied to them, whereby the salience of discourse referents which satisfy parallelism and locality heuristics is boosted, and the salience of discourse referents which the pronoun precedes is decreased. The candidates are then ranked according to salience weight, and the pronoun is interpreted as coreferential with the candidate with the highest salience (or the closest one to it, in the event of a tie). The pronoun is then added to the COREF class, and its salience is recalculated accordingly.

2.4 Evaluation

The anaphora resolution algorithm described here runs at an approximate rate of 75% accuracy, computed as the ratio of correctly resolved pronominal references over all pronouns in text. A totally random selection of text genres, including press releases, product announcements, news stories, magazine articles, and other documents existing as Web pages, has yielded a set of discourses containing 306 pronouns, 231 of which were correctly resolved by our procedure. Details of our evaluation methodology can be found in [5]. We make several observations here.

On its own, this is a respectable rate of accuracy. It is also comparable to that of Lappin and Leass' algorithm, which [6] report as 85%; some deterioration in quality is only to be expected, given the relatively impoverished linguistic base we start with. However, this is not a just a matter of simple comparison, for a number of reasons.

[6] analyzes the output of the procedure applied to a single text genre, computer manuals. Arguably, this is an example of a particularly well behaved text; in any case, it is not clear how the figure would be normalized over a wide range of text types, some of them not completely 'clean', as is the case here. Closer analysis of the current types of error of our algorithm reveals three types of input context which confuse the procedure and contribute to the error rate: occasionally, a pattern for identifying expletives misfires, and the resolution process gets unnecessarily invoked; erroneous pairings of pronouns with antecedents survive all filters, while a simple gender agreement check would rule them out; the resolution algorithm is not sensitive to certain long range contextual (stylistic) phenomena, best exemplified by text containing quoted passages in-line.

Refining the existing pattern for detecting expletives is easily done. Gender (dis-)agreement reflects the lack of gender slot in the LINGSOFT tagger output. It is also possible to adapt the algorithm to take account of e.g. quoted text, given that the input data stream embodies a richer notion of position and context. This suggests that there is room for improvement in the overall quality of our output, bringing it closer in line with that of Lappin and Leass' results.[3]

One final factor needs to be taken into account in this comparison. Lappin and Leass judge their algorithm to have performed correctly on the basis of pairing a pronoun with its correct antecedent. Given the larger goal of our algorithm—anaphora resolution in a wider context—its output is couched in terms of assigning a pronoun to a COREF class. In either case, what current evaluation metrics deem to be a correct output, may still be wrong in a global context. For Lappin and Leass, a pronoun may be correctly paired with an antecedent, but this, in turn, might be incorrectly resolved. For us, assignment of a pronoun to a class might correct, but the class itself might not be

[3] Again, straight comparison would not be trivial, as quoted text passages are not a natural part of computer manuals, and, on the other hand, are an extremely common occurrence in the types of text we are dealing with.

complete. Neither of the evaluation procedures normalizes for such global effects.

3 AN EXAMPLE

Anaphora resolution transforms the logical structure of the text from a set of independent discourse referents to a set of discourse referent COREF classes, in a network of overlaid coreference relations. This additional aspect of our analysis, following from the text segmentation into coherent contexts, is possible after augmenting the algorithm with the CONTEXT-RECENCY salience factor (section 2.2). Ultimately, this enhances the algorithm with the capability to 'track' the salience of a discourse referent across text fragments. Such a record provides valuable information about the text content.

Consider again (see section 2) the text fragment below. The anaphora algorithm is interpreting the emphasized pronoun.

> The group, called the Information Infrastructure Standards Panel (IISP), is sponsored by the American National Standards Institute (ANSI), but is open to all organizations actively working on NII and GII, both members and non-members of ANSI. IISP, which consists of standards developing organizations, industry associations, consortia and architecture groups, companies and government, held *its* first meeting in New York last July.

Following the joint operation of antecedent analysis and prior anaphora resolution, a number of COREF classes are already (partially) instantiated; for instance, the text objects at offsets 176 ("Information Infrastructure Standards Panel"), 181 ("IISP"), and 215 ("IISP"), are collapsed into a single class. "American National Standards Institute" and "ANSI" are likewise conjoined. Similar mechanisms hypothesize, for subsequent mentions of e.g. "the panel", membership to the "IISP" class (`@coref41`). While normally a weak hypothesis, the ability to track salience across larger contexts makes for stronger evidence for ultimately voting conclusively on such a heuristic.

The output of our anaphora resolution algorithm is:[4]

```
Pronoun:        ["its"   [@offset/237]    ]
Candidates:
  @txt_obj162 : ["group" [@offset/162] 180]
  @txt_obj215 : ["IISP"  [@offset/215] 385]
  @txt_obj213 : ["ANSI"  [@offset/213] 155]
  @txt_obj206 : ["NII"   [@offset/215]  34]
  @txt_obj204 : ["GII"   [@offset/215] 140]
```

"Group" has been determined to have a greater weight than "ANSI", even though "ANSI" is referred to more. This is because "group" is a subject, whereas "ANSI" is always in syntactically less-prominent positions. "IISP", on the other hand, is both a subject (in the second sentence) and referred to several times, which correspondingly affects the salience measure of its COREF class. The highest salience weight for this antecedent set determines that "IISP" should be in the same class as "its". In this way not only the pronoun gets locally resolved to its antecedent, but as the antecedent is already identified with a COREF class, other information concerning the class as a whole also becomes associated with the pronoun (for instance, the canonical form of the abbreviated antecedent).

With some simplifications, the final output of the discourse analysis process incorporates the following data fragment:

```
@coref_41 : {
  canform : "information infrastructure
            standards panel"
  trmtype : { PRO ABBREV1 PROPER_NAME ... }
  seglist : { @seg_2[4] @seg_3[1] @seg_4[2] }
  txtobjs : { @txt_obj176 @txt_obj181
              @txt_obj215 @txt_obj237 ... }
            }
```

This indicates that references to a discourse entity identified by a certain canonical form have been found in three (out of total five) segments in the text, in a variety of textual forms (including a pronominal reference and an abbreviation). Four occurrences alone have been observed in the second segment (`@seg_2[4]`), in which the example fragment earlier occurs.

4 CONCLUSION

We have looked at anaphora resolution as an integral component of a discourse processing model concerned with building a richer representation of logical structure of text. Moreover, our approach seeks to establish and maintain coreferentiality among all entities in the discourse, *without* relying upon full syntactic analysis of the input. To this end, we have succesfully adapted Lappin and Leass' algorithm, without compromising overall quality. Our strategy for circumventing the need for full syntactic parse is applicable to other interpretation tasks which, similarly to anaphora resolution, belong to higher level semantics and discourse.

The algorithm we present generates a record of the overall salience of discourse referents as a function of the individual salience of each member of a coreference class. The two complementary notions discussed here—salience computed across an entire text and incrementally instantiated coreference classes—are powerful devices for content-rich discourse processing.

REFERENCES

[1] Marti Hearst, 'Multi-paragraph segmentation of expository text', in *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, (June 1994).

[2] Michael Johnston, Branimir Boguraev, and James Pustejovsky, 'The acquisition and interpretation of complex nominals', in AAAI Workshop on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity, Stanford, California, (March 1995).

[3] John Justeson and Slava Katz, 'Technical terminology: some linguistic properties and an algorithm for identification in text', *Natural Language Engineering*, **1**(1), 9–27, (1995).

[4] Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Antilla, *Constraint grammar: A language-independent system for parsing free text*, Mouton de Gruyter, Berlin / New York, 1995.

[5] Christopher Kennedy and Branimir Boguraev, 'Anaphora for everyone: Pronominal anaphora resolution without a parser', in *Proceedings of COLING-96*, Copenhagen, DK, (1996).

[6] Shalom Lappin and Herb Leass, 'An algorithm for pronominal anaphora resolution', *Computational Linguistics*, **20**(4), 535–561, (1994).

[7] Inderjeet Mani and T. Richard MacMillan, 'Identifying unknownproper namesin newswiretext', in *Corpus Processing for Lexical Acquisition*, eds., Branimir Boguraev and James Pustejovsky, 41–60, MIT Press, (1996).

[8] Woojin Paik, Elizabeth Liddy, and Edmund Yuand Mary McKenna, 'Categorizing and standardizing proper nouns for effcient information retrieval', in *Corpus Processing for Lexical Acquisition*, eds., Branimir Boguraev and James Pustejovsky, 61–74, MIT Press, (1996).

[9] Atro Voutilainen, Juha Heikkilä, and Arto Antilla. A constraint grammar of english: A performance-oriented approach. University of Helsinki, Department of General Linguistics, Publication No. 21, Hallituskatu 11–13, SF-00100 Helsinki, Finland, 1992.

[4] Note that our syntactic filters are quite capable of discarding a number of configurationally inappropriate antecedents, which appear to satisfy the precedence relation.