Measuring the Influence of Errors Induced by the Presence of Dialogues in Reference Clustering of Narrative Text

Alaukik Aggarwal
Dept. of Computer Science and Eng.
Maharaja Agrasen Institute of Technology
Delhi, India
alaukik.aggarwal@gmail.com

Pablo Gervás, Raquel Hervás
Instituto de Tecnología del Conocimiento
Universidad Complutense de Madrid
Madrid, Spain
pgervas@sip.ucm.es
raquelhb@fdi.ucm.es

Abstract

The task of identifying noun phrases in a text that co-refer to the same discourse entity is of fundamental importance in natural language understanding. An important, often overlooked, challenge in this task is the difficulty posed by noun references appearing within quoted fragments of dialogue, due to the warping of the context introduced by the quotation. This paper presents a quantitative analysis of the effect of this problem in the performance of a clustering approach to coreference resolution. A coreference resolution algorithm based on clustering is applied to a set of texts that contain dialogues and to a modified version of the texts where the dialogue quotes have been eliminated. The results of precision and coverage over both sets are compared. The differences in performance constitute a significant indication of the relevance of the studied problem for the kind of texts under consideration.

1 Introduction

The task of identifying noun phrases in a text that co-refer to the same discourse entity is of fundamental importance in natural language understanding. This task has been addressed in various research contexts such as information extraction (Kameyama, 1997), question answering (Morton, 2000), and automatic summarization (Azzam et al., 1999). For these tasks, reasonable results are reported. An important, often overlooked, challenge in this task is the difficulty posed by pronominal references appearing within quoted fragments of dialogue, where the referential value of personal pronouns is significantly mediated by the roles of speaker and addressee (and the set of listeners present in the dialogue context), and the referential value of demonstratives may change dramatically depending on the particular speaker in a given dialogue turn. In general, most attempts at coreference resolution have focused on texts of specific genres (news articles, questions, simple instructional or informational dialogues) where no nested dialogues appear as direct speech. However, a large amount of text as processed by humans, particularly narrative text, tends to include a significant percentage of dialogue in quoted form. With a view to extrapolating the applicability of the coreference resolution algorithms already developed to this large body of textual material, it would be interesting to obtain a quantitative measure of the effect of quoted dialogue in the performance results. This is the goal of the present paper.

This paper presents a quantitative analysis of the effect of this problem in the performance of a clustering approach to coreference resolution. A coreference resolution algorithm based on clustering is applied to a set of texts that contain dialogues, and to a modified version of the texts where the dialogue quotes have been eliminated. The results of precision and coverage over both sets are compared. The differences in performance constitute a significant indication of the relevance of the studied problem for the kind of texts under consideration.

Section 2 provides a brief overview of existing work on coreference resolution, outlines the problem of references in dialogue, and describes the auxiliary NLP tools employed in the paper. Section 3 describes the particular corpus employed for the tests. Section 4 describes the algorithm employed. Section 5 presents the evaluation of the effect of the presence of dialogues over the two versions of the corpus, and the two final sections discuss the results and outline our conclusions.

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

2 Previous Work

Coreference resolution involves, given a text, identifying which noun phrases in the text refer to the same entity. Noun phrases can be a definite or indefinite noun phrase, a pronoun, a demonstrative, or a reflexive. But proper nouns, and what are known as named entities, may also be the subject of references. Coreference resolution usually establishes the particular entities referred by each reference in the text. It has been an exhaustively studied problem in Natural Language Processing.

2.1 Approaches to Coreference Resolution

Coreference resolution has been attempted both using knowledge-rich and knowledge-poor approaches. Knowledge-rich approaches generally relied on hand-crafted resources and hand-crafted input (carefully preprocessed sets of elements, rather than actual text), and focused largely on resolution of pronouns. Knowledge-poor approaches relied on machine learning techniques and fully automated processing of the input from the source text, and generally deal with a broader range of noun phrases. Classic examples of knowledge-rich approaches are (Lappin and Leas, 1994) and (Mitkov, 1997). For the present paper, we are concerned with knowledge-poor approaches, but we will address in the discussion the possible relevance of knowledge-rich solutions for the particular case of dialogue.

Decision trees were used for coreference resolution in (Mccarthy and Lehnert, 1995). The authors presented RESOLVE, a system that used decision trees in order to learn how to classify coreferent phrases in the domain of business. The performance of RESOLVE was compared to the performance of a manually generated set of rules for the same task. The results showed that higher performance was obtained using decision trees for two of the three evaluation metrics they tested. Although the obtained were quite high, the approach had the problem of requiring a corpus of annotated texts to create the decision trees.

Another learning approach to coreference resolution was explored in (Soon et al., 2001), where an updated version of C4.5 learning algorithm was used to build classifiers that dealt with this problem. The learning algorithm required a relatively small corpus of training documents annotated with coreference chains of noun phrases. Their system was the first learning-based system that offered results comparable to that of other non-learning systems of literature.

In (Cardie and Wagstaff, 1999) the authors presented the idea of using clustering algorithms to deal with the problem of coreference resolution. In their algorithm, coreference between two noun phrases was determined on the basis of some list of attributes called *feature vector*. First, they computed the feature vector for each noun phrase and assigned weights to each of them. Then they applied the clustering algorithm, that basically calculated the conceptual *distance* between noun phrases and grouped the ones that a have a distance lower than a certain radius *r*. The conceptual distance between two noun phrases under consideration was defined as in Equation 1.

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f * incmp_f(NP_i, NP_j) \quad (1)$$

where F represents the feature vector, NP_j is the current noun phrase, and NP_i is each noun phrase in the input text occurring before NP_j. $incmp_f$ represents an incompatibility function that finds whether the two noun phrases are related to each other or not. w_f assigns a weight to each feature in the feature vector representing the importance of the feature in finding a coreference.

The value r represents the clustering radius threshold, so NP_i and NP_j corefer if $dist(NP_i, NP_j) < r$. Terms with a weight of ∞ represent that two noun phrases can never corefer when they have incompatible values for that feature. Terms with a weight of $-\infty$ force coreference between two noun phrases with compatible values for that feature.

Several possible improvements on machine learning solutions for coreference resolution are discussed in (Ng and Cardie, 2002). They propose an extension of the set of features to enrich the knowledge set available to the algorithm used in (Soon et al., 2001). They observe that extending the set of features results in immediate performance drops. Though these can be remedied with manual selection of features, the result of selection performs well

on common nouns but badly on pronouns. They also propose three extra-linguistic modification to the original Soon Algorithm which led to significant improvement in performance.

A number of hybrid approaches have emerged recently. One such is (Bergsma and Lin, 2006), which applied a Support Vector Machine pronoun resolution classifier over knowledge-rich dependency trees to learn coreferent and non-coreferent dependency paths between the two entities as they appear in the syntax tree.

2.2 Pronominal Reference in Dialogues

The problems inherent to dealing with pronominal references in the context of a discourse involving dialogues are described in detail in (Callaway and Lester, 2002). Their main argument is illustrated in examples such as:

(1) "I think I will go find my shoes", said John.
(2) "You should go find your shoes", said John.
(3) "We should go find your shoes", said John.
(4) "They went to eat their breakfast", said John.

In such cases, it may be possible to identify the pronoun *I* as a reference to John in (1). But it is impossible to identify the correct referents for the pronouns *you*, *we* and *they* unless the dialogue situation is taken into account. For instance, if (2) (3) and (4) involve statements addressed by John to Mary, *you* will be assigned to Mary and *we* will be assigned to the set composed of John and Mary. To resolve the reference for *they* it will be important to take into account that neither John nor Mary are intended as referents.

The difficulties arise from the fact that the same surface form for the pronouns will refer to different entities depending on the dialog context in which it occurs. Appropriate treatment of this problem would require a certain level of identification of dialogue structure, which is currently not contemplated by approaches to coreference resolution. The present paper attempts to qualify the impact of this oversight.

Similar problems have also been highlighted in other studies. In (Strube and Müller, 2003) the authors suggested that most of the NPs found in dialogs have non-NP antecedents or no antecedents at all (like in *It is raining*). They based their findings in previous works where in different corpora it was found that about 50% of the pronouns have non-NP-antecedents. (Byron, 2002) presented a symbolic approach which resolved pronouns with NP and non-NP antecedents in spoken dialogue in the TRAINS domain. In (Delmonte, 2002) the problem of subject NP being empty or NP in post-verbal position (with pre-verbal NP missing) was also addressed. The author studied different features in different languages based on which he defined some rules to identify and differentiate between direct speech and other kind of utterances. It is also worth pointing out the work in (Navarretta, 2004), where the author presents an approach based on salience in the hearer's cognitive model for resolving intersentential pronominal anaphora in Danish texts and dialogues.

2.3 Linguistic Resources Used for Preprocessing Texts

In the field of natural language analysis and understanding, GATE (General Architecture for Text Engineering) (Cunningham et al., 1995) is an architecture that provides an infrastructure for building language engineering (LE) systems or a development environment to aid construction, testing and evaluation of LE systems. GATE goes beyond the definition of interfaces and standard data structures to provide a set of resources known as CREOLE (a Collection of REusable Objects for Language Engineering).

WordNet (Miller, 1995) is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. The most ambitious feature of WordNet is its attempt to organize lexical information in terms of word meanings, rather than word forms. English nouns, verbs and adjectives are organized into synonyms sets, each of them representing one underlying lexical concept. These synonyms sets - or *synsets* - are linked by semantic relations like synonymy or hyponymy.

Its organization by concepts rather than word forms allows WordNet to be used also like a knowledge source. The hyponymy/hypernymy relation can be considered equivalent to the "isa" one, and all the concepts are hyponyms of a reduce set of general

concepts that form the conceptual root of WordNet.

As a resource for obtaining gender data, we have used (Bergsma and Lin, 2006). In their gender data they have three values corresponding to each word. First value corresponds to MALE, second to FEMALE and third to NEUTRAL. Each value for the word corresponds to the number of times the word has been used in some text in that particular gender form.

3 The Corpus of Narrative Texts

As corpus we have selected 30 folk tales with different sizes and all of them written in English. The idea was to cover different styles by having tales from different authors and time periods. The reason for using folk tales is that they usually contain many characters and objects involved in the story, so they are rich in coreferences. We have also looked for tales with dialogs between the characters in order to test the difficulties of coreference resolution in these cases.

In order to execute a coreference resolution algorithm to identify coreference chains from a text, it is required to obtain some extra information from the raw text about sentences, words and other linguistic information. Specially, it is necessary to identify the references from the text that would be used in the coreference algorithm, and to extract the feature vectors for each reference. For the work presented in this paper we have decided to use GATE to enrich the initial text with such information.

3.1 Identifying References Using GATE

We have considered as references all the noun phrases that are present in the text and we have used GATE to obtain these noun phrases. GATE provides a graphical interface where a corpus of text can be easily created and processed. In this case we have composed a processing pipeline (using different modules provided by the tool) with the following steps:

1. The text is divided into sentences by the ANNIE Sentence Splitter.
2. The text is split into tokens by using the ANNIE English Tokeniser.
3. For each token, its part of speech (POS) is determined by using the ANNIE POS Tagger.
4. The noun phrases are identified by using the Noun Phrase Chunker CREOLE plugin.

The resulting document can be stored in XML format so it can be read easily. This output XML file is divided into three main parts.

The first one contains the whole text with each space between tokens marked with a numbered node. These nodes are used to delimitate the elements that are enumerated in the rest of the file.

The second part of the file contains the information about all the tokens that have been identified by the tokeniser. These tokens can be text chunks, spaces, sentences and noun phrases. In Figure 1 the information given for a token is shown. This information goes from the start and end position where the token is appearing, to the POS or the category corresponding to the token. It is important to point that tokens are not only words, but also spaces and punctuations marks, between others.

```
<Annotation Id="301" Type="Token" StartNode="634" EndNode="642">
    <Feature>
        <Name className="java.lang.String">length</Name>
        <Value className="java.lang.String">8</Value>
    </Feature>
    <Feature>
        <Name className="java.lang.String">category</Name>
        <Value className="java.lang.String">NN</Value>
    </Feature>
    <Feature>
        <Name className="java.lang.String">orth</Name>
        <Value className="java.lang.String">lowercase</Value>
    </Feature>
    <Feature>
        <Name className="java.lang.String">kind</Name>
        <Value className="java.lang.String">word</Value>
    </Feature>
    <Feature>
        <Name className="java.lang.String">string</Name>
        <Value className="java.lang.String">merchant</Value>
    </Feature>
</Annotation>
```

Figure 1: Information about the token *merchant*

The third part of the file contains the list of noun phrases found in the text. They are also identified with their position in the text using the start and end node.

3.2 Feature Extraction

The XML files obtained after processing the raw text using the GATE tool contained a lot of extra information. We extracted the noun phrase information from GATE, the string of text occurring in the text,

part-of-speech[1], sentence and paragraph order, and the order of its occurrence using the `startNode` and the `endNode`. The extracted information from GATE output was combined with information from other resources to form a second XML file.

This second XML file, that served as input to our algorithm, corresponds to a representation of the whole text where the noun phrases, as identified by GATE, are XML nodes called `reference`. Inside each noun phrase, nouns and pronouns were identified and separate nodes (called `nucleus`) were created inside the `reference` elements. This nucleus is very useful to achieve higher accuracy in identification. Information related to the `nucleus` node is added in form of attributes.

The attributes applied to `reference` and `nucleus` nodes are the following:

Position. Its value is assigned on the basis of `startNode` and `endNode` attributes from the GATE file. References are sequentially numbered from the beginning of the document. This attribute is used in the `reference` nodes and not in the `nucleus`.

POS. Part Of Speech was derived directly from the GATE output. As we have added the information only for the `nucleus`, we identified three specific values for this. The values could be NOUN, PRONOUN or PROPERNOUN.

Article. If the text preceding the `nucleus` node is either *a* or *an*, then its value is assigned to be INDEF. If the text is *the*, then its value is DEF. In all other cases, its value is NONE.

Number. Its value is determined separately for nouns and pronouns. For nouns, GATE directly identifies whether the noun phrase is singular or plural. But for pronouns, we separately identified a list of pronouns and their number. The number can be SING (singular), PLURAL or ANY. We use ANY to denote those nouns or pronouns that can be both SING and PLURAL, like *who* can be used to refer both singular and plural nouns.

An special case to take into account about number are the references in the form of a conjunction. We found that sometimes we had plural references that were a conjunction of two singular references that appeared separately in the text. An example would be *Marigold and Dressalinda*, where Marigold and Dressalinda are names of girls and would be singular when treated as individual and plural when together. Hence, we have added some extra information about such nodes inside the `reference` nodes representing them by `copy` nodes. In the algorithm we look for coreference for both the `copy` and the `nucleus` inside the `reference`. This help us to identify the coreference between those noun phrases separately or jointly, allowing us to find coreferences with both singular (*she*) or plural (*they*) pronouns.

Semantic Class. WordNet returns a hypernym tree corresponding to each word searched. From this tree, we extracted the semantic classes corresponding to the lexical entries for nouns in the text. Out of the large number of semantic classes available in WordNet we used HUMAN (for person, relation or group returned from the semantic class), NON_HUMAN (for animal or plant), OBJECT (artifact, communication, food, object, substance or possession), BODY, LOCATION, TIME, STATE, COGNITION, ATTRIBUTE (attribute or shape), MOTIVE, PHENOMENON (phenomenon or process), QUANTITY, FEELING, EVENT, ACT and TOPS.

We created files corresponding to different semantic classes used for pronouns. We identified that the semantic class for a pronoun can either be HUMAN or ANY. Semantic class ANY was used to correspond to pronouns that can be used to refer both HUMAN and NON_HUMAN nouns, like for the case of interrogative pronoun *what*.

Gender. We identified the gender a word using the works of (Bergsma and Lin, 2006). We simply use the gender that has highest count corresponding to the word.

We manually identified and created files cor-

[1]Part-of-speech tags used in GATE are available at http://gate.ac.uk/sale/tao/index.html#x1-449000D

```
<reference position="28" string="the merchant">
   the
   <nucleus pos="NOUN" article="DEF" number="SING"  semClass="HUMAN" gender="MALE" id="1" >
      merchant
   </nucleus>
</reference>


<reference position="18" string="Marigold and Dressalinda">
   <nucleus pos="PROPERNOUN" article="NONE" number="SING"  semClass="NON_HUMAN" gender="FEMALE" id="1">
      Marigold
   </nucleus>
   and
   <nucleus pos="PROPERNOUN" article="NONE" number="SING"  semClass="NOT_FOUND" gender="NOT_FOUND" id="2">
      Dressalinda
   </nucleus>
   <copy pos="PROPERNOUN" article="NONE" number="PLURAL" semClass="NOT_FOUND" gender="UNKNOWN" id="1" >
      Marigold and Dressalinda
   </copy>
</reference>


<reference position="36" string="he">
   <nucleus  pos="PRONOUN" article="NONE" number="SING" semClass="HUMAN" gender="MALE" id="1">
      he
   </nucleus>
</reference>
```

Figure 2: Some examples of the enriched references in XML

responding to the pronouns labeled as either MALE, FEMALE or NEUTRAL.

Examples of the XML annotation of some references in the text can be seen in Figure 2. Each example corresponds to a different kind of reference (noun, plural proper noun or pronoun), showing the different attributes that are associated to each references as explained above.

4 A Clustering Algorithm for Coreference Resolution

We have explored the coreference resolution as a clustering problem, basing our algorithm in (Cardie and Wagstaff, 1999). Note that, as described in section 2.1, this kind of algorithm was based in a feature vector assigned to each noun phrase that is a candidate for coreferring. Using this feature vector, a conceptual distance between two noun phrases can be calculated. If this distance is lower than a certain defined radius, then the noun phrases are coreferring.

4.1 Feature Vector

For determining the coreference between the noun phrases, we need to devise a list of features required to find the associations between noun phrases. We used more or less the same feature vector than in the original algorithm (Cardie and Wagstaff, 1999), but keeping in mind that our corpus consists of tales. So, there are some kinds of noun phrases that we would not encounter like the names of companies (usually found in news articles) and designations of people in companies (marketing officer, chairman, president etc).

Our feature vector consists of seven features (instead of the original eleven). They work on the basis of two extracted noun phrases, NP_i and NP_j, where NP_i is the antecedent noun phrase and NP_j is the current noun phrase under consideration.

Position. This feature is used to calculate the distance between two noun phrases taking into account their positions in the text. Integer values are assigned to each noun phrase found in the text, starting from the beginning.

Pronoun. Its value can be determined from the `POS` attribute used in our input XML file. If the value of `POS` attribute is PRONOUN, then it is simply recognized as a pronoun.

Article. This was used to identify noun phrases that were preceded by indefinite or definite articles. Their possible values are the same than the ones for the `article` node: INDEF (indefinite, contains *a* or *an*), DEF (definite, contains *the*) or NONE (in other cases). For example, *the merchant* is a definite noun phrase.

Word-Substring. This feature is used to check if the nucleus of the noun phrase NP_j is present

Feature	Weight	Incompatibility
Position	5.0	Difference in value of position for each noun phrase
Pronoun	r	1 if NP_i is a pronoun and NP_j is not; else 0
Article	r	1 if NP_j is indefinite; else 0
Word-substring	$-\infty$	1 if NP_i includes NP_j entirely as a substring; else 0
Number	∞	1 if they do not match in number; else 0
Semantic class	∞	1 if they do not match in semantic class; else 0
Gender	∞	1 if they do not match in gender; else 0

Table 1: Incompatibility functions and weights for the elements in the feature vector

as a substring in the noun phrase NP_i. For example, if the nucleus in noun phrase NP_j is *merchant* and the noun phrase NP_i is *a rich merchant*, the nucleus of noun phrase NP_j is present as a substring in noun phrase NP_i.

Number. If the number of both the nucleus from noun phrase NP_i and noun phrase NP_j matches, then we say that the two nucleus match in number. The value of number might be SING (singular), PLURAL or ANY.

Semantic Class. Semantic class is a way to compare the basic sense of the word. If the semantic class of the nuclei of two noun phrases NP_i and NP_j match, then we can say that the two nucleus match in the semantic class. The value of semantic class can be any of the possible values of the `SemanticClass` element. This feature was one of the most important. As we were able to identify many different semantic classes, we were able to differentiate between noun phrases more accurately.

Gender. The value for this feature can be MALE, FEMALE or NEUTRAL. If the nucleus of both noun phrases NP_i and NP_j match, then they can be said to be matching in gender. Also, if the gender of either of the noun phrase is NEUTRAL then it can match with the other noun phrase, irrespective of its value for the gender.

4.2 Distance Between Noun Phrases

A distance matrix is used to compute the conceptual distance between two noun phrases under consideration, using the equation shown in 2.1. For our algorithm we have redefined the incompatibility function and weights for each feature. They are displayed in Table 1.

5 Evaluation

In order to evaluate the influence of dialogs when dealing with the problem of coreference resolution, we used the described clustering algorithm over two different set of texts: with and without dialogs. The whole corpus of tales was used for both sets, but for the second one the tales were processed automatically to eliminate all kind of dialog. Note that in this automatic processing we only eliminated the content of the dialog, but not the rest of the direct speech. For example, the piece of text *"You are lying!", she said* would be changed to *"", she said* after this processing.

For the comparison between the obtained coreference clusters and the correct ones, we hand-inspected the tales and identified the right coreferring noun phrases. Then we used a simple comparison algorithm to obtain the precision and recall of our algorithm.

Table 2 summarizes the obtained results for both tales with and without dialogs. The difference between considering the dialogs or not when calculating the coreferences in a text is of a 10%, that can be considered quite high. This reinforces the idea about how the dialogs in a text must be considered in a different way when dealing with coreference resolution.

Those results where obtained with radius $r = 31$. We tested the algorithm with different radius values in order to find the one that obtained higher precision and recall results. Table 3 shows the results obtained with different r values.

	Precision	Recall
With dialogs	61.10	56.57
Without dialogs	70.49	63.15

Table 2: Precision and recall results for both texts with and without dialogs

Radius	With dialogs		Without dialogs	
	Recall	Precision	Recall	Precision
10	36.81	50.93	41.95	62.69
15	43.52	53.75	49.61	64.97
20	47.97	55.86	53.78	66.77
25	51.41	57.29	54.44	66.11
30	53.77	59.26	57.01	66.79
31	56.57	61.10	63.15	70.49
36	57.06	61.26	60.63	67.35
37	56.76	60.76	60.07	66.37
40	56.76	60.76	60.07	66.37
50	56.54	59.77	60.55	65.84
58	55.62	58.81	59.80	64.23
67	55.19	58.39	58.63	62.96

Table 3: Precision and recall results for different r values

6 Discussion

The ablation of all quoted fragments of dialogue results in a simultaneous 9 % improvement in precision and 7 % improvement in recall. This suggests that the effect of errors originating in dialogue context on the overall performance of coreference resolution approaches to narrative text can be quite significant. These number may become considerably higher if the percentage of dialogue in the texts increases.

It is clear that ablation is no practical solution, as more information is lost in removing the dialogue than may have been lost due to errors in coreference annotation. However, these results indicate that the task of solving coreference in dialogue contexts is worth addressing. Even if initial solutions perform badly in relation with the total volume of dialogue, they may improve overall performance significantly in texts with a large proportion of dialogue.

A reasonable way to achieve this may be to extend the set of features taken into account by the algorithm to include information about the dialogue context (whether a conversation is being reported, who is the speaker, who is the addressee). These would have to be integrated with the existing set of features. The clustering solution selected in the paper provides a reasonable balance between complexity and performance. In view of the mixed results obtained by (Ng and Cardie, 2002) in extending the set of features used by the algorithm, we have preferred to retain a smaller initial set of features. Further work may address the different possibilities of improving the algorithm detailed by Ng and Cardie.

It remains to be seen whether the information required to implement features describing the dialogue situation might be easily obtained from existing NLP solutions available for preprocessing. Mitkov (Mitkov et al., 2001) argues that inaccuracies in the preprocessing stage may account for a large percentage of the accumulated error in systems for coreference resolution that use automated preprocessing, and (Bergsma and Lin, 2006) observe that direct comparison between different approaches is made difficult by the differences in preprocessing. The approach followed in this paper for preprocessing has been described in detail so that possible sources of error or differentiation with other solutions may be identified.

If the task of obtaining the necessary dialogue features through automatic preprocessing proves fruitless, an attempt could be made to develop a solution based on hand-crafted input. This may be more in line with the knowledge-rich approaches to coreference resolution mentioned above.

With respect to the issue of word sense disambiguation, a trivial solution of taking the first sense available for each word has been applied. More refined solution, such as the one applied in (Muñoz et al., 2002) could be applied.

Because the approach presented here considers a wide range of noun phrases, it is actually addressing simultaneously tasks that have been addressed in specific research efforts elsewhere, such as named entity recognition for identifying proper nouns (McCallum and Wellner, 2003), coreference of definite descriptions (Vieira and Poesio, 2000), and more specific solution for anaphora resolution for the specific case of third person pronouns described above. It is clear that solutions tailored to a narrower range of phenomena are bound to obtain better performance for their specific target elements, the goal of

this paper was to measure the effects of not taking into account the challenges of personal pronouns in contexts of dialogue on the overall performance of coreference resolution over narrative texts. For this reason, a solution with the widest possible coverage has been selected.

Nevertheless, some advantages of the approach presented in this paper are:

- The clustering approach is unsupervised, so an annotated corpus is not required for training the algorithm. In our case this is a great advantage as there is no available corpus of tales annotated with coreferences.
- Although we have applied the algorithm to the story domain, the clustering approach is domain-independent. It will work in the same way in any other kind of texts, both with and without dialogs.
- The clustering approach permits to coordinate context-independent and context-dependent constraints and preferences for partitioning noun phrases into coreference clusters.

Some words must be said concerning the relative importance of dialogue in narrative texts, and the relative importance of narrative texts in human experience. For narrative texts, fragments presented in the form of dialogue can be a fundamental part of the total conveyed information. Most stories, from the simplest fairy tale to long novels, include parts where conversations between two or more characters are reported. And yet there is no mention of the issue of dialogue as a possible feature to take into account in any of the existing systems for coreference resolution. This may be due in part to the difficulties presented by the automated processing of dialogue. Whereas many existing parsers achieve very high performance scores in simple texts, the dialogue formats tend to confuse them, resulting in jumbled parse trees. In the absence of data, this may be attributed to the fact that most parsers are trained on corpora that include very little dialogue. However, the issues may run deeper than that, and it is possible that specific solutions may be required for automatically handling dialogue situations.

On a similar tack, narrative texts constitute a very significant portion of human culture. To date, there has been very little effort to apply natural language processing techniques to this treasure trove of texts. This may in itself justify the absence of work on such fundamental issues. However, in the long run this is a problem that needs to be solved.

Existing work on anaphora resolution has tended to sidestep the issue of first and second person personal pronouns. In (Lappin and Leas, 1994) and (Morton, 2000), the range of pronouns is explicitly restricted to third person pronouns. Texts used for training and testing normally arise from existing annotated corpora, which very rarely include literary or narrative material. A notable exception is (de Arruda Santos and Carvalho, 2007), which test a version of Hobb's algorithm for pronoun resolution in Portuguese over three different corpora. One of them is a literary corpus. Unfortunately, no data is given in the paper as to whether the literary corpus (or any one of the others) contained any significant percentage of dialogue. Nevertheless, it may be significant that a a decrease of nearly 10% is observed in the success rate between the magazine corpus and the literary corpus.

7 Conclusions and Future Work

The results obtained in this paper indicate that overlooking the challenges of reference of personal pronouns in dialogue contexts may result in performance degradations over an average story verging on 10%. These numbers may increase for texts with larger proportions of dialogue. This suggest that the task of coreference resolution for dialogues is worth addressing if large volumes of narrative text are likely to be processed in the future.

Such an endeavor faces two basic difficulties at the level of automatic preprocessing of text. Current parsers perform badly on the task of extracting correct syntactic structures from sentences involving quoted speech. This problem would have to be solved before automatic preprocessing of dialogues can be attempted with any hope of success. Additionally, it is unclear whether the information that might be useful for coreference resolution in dialogue could be extracted by simple syntactic processes, as it will generally involve pragmatic information.

In view of these conclusions, it seems probable

that knowledge-rich solutions to coreference resolution in dialogue, operating over hand-crafted input that includes the required pragmatic information, may have a better chance of succeeding in the near future than fully automated machine learning techniques.

Acknowledgments

This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project).

References

de Arruda Santos, D.N. and Carvalho, A.M.B.R. 2007. Hobbs Algorithm for Pronoun Resolution in Portuguese . In *MICAI 2007*, LNAI 4827, Springer, pages 966–974.

Azzam, S. and Humphreys. K. and Gaizauskas, R. 1999. Using coreference chains for text summarization. In *Proceedings of the Workshop on Coreference and Its Applications*, College Park, Maryland, June.

Bergsma, S. and D. Lin. 2006. Bootstrapping Path-Based Pronoun Resolution. In *Proceedings of the Conference on Computational Lingustics*, Association for Computational Linguistics (COLING/ACL-06), Sydney, Australia.

Byron, D.K. 2002. Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Penn., 7-12 July 2002, pp. 80-87.

Callaway, Charles B. and Lester, James C. 2002. Pronominalization in generated discourse and dialogue. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 88–95.

Cardie, C. and K. Wagstaff. 1999. Noun Phrase Coreference as Clustering. In *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora*, pages 82–89.

Cunningham, H., R.G. Gaizauskas, and Y. Wilks 1995. A General Architecture for Text Engineering (GATE) - a new approach to Language Engineering. *Technical Report CS-95-21*, Department of Computer Science, University of Sheffield.

Delmonte, R. 2002. GETARUN PARSER: A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG. In *Proceedings of the LFG-02 Conference*, Athens, pages 130–153.

Kameyama, M. 1997 Recognizing Referential Links: An Information Extraction Perspective. In *Proceedings of the ACL '97/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*.

Lappin, S. and Leas, H. 1994 An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20 (4), 535–562.

Miller, G.A. 1995 Wordnet: a lexical database for English. *Communications of the ACM*, 38 (11), 39–41.

Mitkov, R. 1997 Factors in Anaphora Resolution - A case study on two different approaches. In *Proceedings of the ACL '97/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*.

Mitkov, R. and Bougarev, B. and Lappin, S. 2001 Introduction to the Special Issue on Computational Anaphora Resolution. *Computational Linguistics*, 27 (4).

Morton, S. 2000. Coreference for NLP applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 173–180.

Muñoz, R. and Saiz-Noeda, M. and Montoyo, A. 2002. Semantic Information in Anaphora Resolution. In *Proceedings of the Third International Conference on Advances in Natural Language Processing*.

McCallum, A. and Wellner, B. 2003. Toward Conditional Models of Identity Uncertainty with Application to Proper Noun Coreference. In *NIPS*, MIT Press, pages 905–912.

McCarthy, J.F. and Lehnert, W.G. 1995. Using Decision Trees for Coreference Resolution. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1050–1055.

Navarreta, C. 2004. Resolving Individual and Abstract Anaphora in Texts and Dialogues. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, Geneva, Switzerland, pages 233–239.

Ng, V. and Cardie, C. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–11.

Soon, W.M. and Ng, H.T. and Daniel, C.Y.L. 2001 A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27 (4), 521–544.

Strube, Mi. and Müller, C. 2003. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 03)*, pages 168–175.

Vieira, R. and Poesio, M. 2000 An empirically based system for processing definite descriptions. *Computational Linguistics*, 26 (4), 539–593.