Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries

Advaith Siddharthan*
University of Aberdeen

Ani Nenkova**
University of Pennsylvania

Kathleen McKeown†
Columbia University

Although there has been much theoretical work on using various information status distinctions to explain the form of references in written text, there have been few studies that attempt to automatically learn these distinctions for generating references in the context of computer-regenerated text. In this article, we present a model for generating references to people in news summaries that incorporates insights from both theory and a corpus analysis of human written summaries. In particular, our model captures how two properties of a person referred to in the summary—familiarity to the reader and global salience in the news story—affect the content and form of the initial reference to that person in a summary. We demonstrate that these two distinctions can be learned from a typical input for multi-document summarization and that they can be used to make regeneration decisions that improve the quality of extractive summaries.

1. Introduction

News reports, and consequently news summaries, contain frequent references to the people who participate in the reported events. Generating referring expressions to people in news summaries is a complex task, however, especially in a multi-document summarization setting where different documents can refer to the same person in different ways. One issue is that the generator has to work with textual input as opposed to closed-domain semantic representations. More importantly, generating references to people involves issues not generally considered in the referring expression literature.

* Department of Computing Science, University of Aberdeen, Meston Building, Meston Walk, Aberdeen AB24 3UE, UK. E-mail: advaith@abdn.ac.uk.
** University of Pennsylvania, CIS, 3330 Walnut St., Philadelphia, PA 19104, US. E-mail: nenkova@seas.upenn.edu.
† Department of Computer Science, Columbia University, 1214 Amsterdam Ave., New York, NY 10027, US. E-mail: kathy@cs.columbia.edu.

Submission received: 30 May 2008; revised submission received: 18 July 2010; accepted for publication: 25 March 2011.

People have names that usually distinguish them from others in context, and as such, the framework of generating descriptions that rule out distractors (e.g., the body of research building on Dale [1992], including the recent shared tasks based on the TUNA corpus [Gatt, Belz, and Kow 2008]) is not appropriate. Recent GREC challenges (Belz, Kow, and Viethen 2009) have, however, focused on references to named entities, and we compare that body of work to ours in Section 2.2.

In our domain of summarizing news articles, the aim of generating references to people is to introduce them to the listener and relate them to the story. How much description is needed depends on many factors, including the knowledge the speaker expects the listener to have, the context of the discourse, and the importance of the person to the narrative. In this article, we explore these three information status distinctions in the context of generating references to people in multi-document summaries of newswire:

1. **discourse-new vs. discourse-old**: whether a reference to a person is a first or subsequent mention of that person is purely a property of the text.
2. **hearer-new vs. hearer-old**: whether the person being mentioned is familiar to the hearer is a property associated with the hearer.
3. **major vs. minor character**: how important or salient a person is to the narrative depends on communicative goals and is therefore a property associated with the speaker.

Through our studies, we seek answers to three main research questions:

1. Is it possible to automatically infer information not explicitly stated about people in the input for summarization, such as familiarity and salience?
2. Is such information useful for the task of generating references to people in multi-document summaries?
3. Can summary quality be improved through an informed rewrite of references to people?

We report positive answers to all three questions. Our corpus-based approach models the differences between first and subsequent references, provides detail on how to generate the variety of first references that occur, and shows how distinctions such as familiarity and salience drive generation decisions for initial references.

In this article, we describe and evaluate a new algorithm for referring to people in multi-document news summaries that integrates results from two earlier studies (Nenkova and McKeown 2003; Nenkova, Siddharthan, and McKeown 2005). In the following sections, we first discuss related work (Section 2) and then present a simple model for distinguishing discourse-new and discourse-old references (Section 3, based on Nenkova and McKeown [2003]). A more sophisticated model based on automatically acquired information about familiarity and salience is presented in Section 4 and these distinctions are used for making generation decisions in Section 5 (based on Nenkova, Siddharthan, and McKeown [2005]). We then present an integrated algorithm that generates new references to people in news summaries based on the acquired information status distinctions and report an evaluation of the effect of reference rewriting on summary quality in Section 6, including a discussion of its scope and limitations in Section 6.2.

2. Related Work

Related research into summarization, information status distinctions, and generating referring expressions is reviewed here.

2.1 Extractive and Abstractive Summarization

Multi-document summarization has been an active area of research over the past two decades and yet, barring a few exceptions (Radev and McKeown 1998; Daumé III et al. 2002; Barzilay and McKeown 2005), most systems still use shallow features to produce an extractive summary, an age-old technique (Luhn 1958) that has well-known problems. Extractive summaries may contain phrases that the reader cannot understand out of context (Paice 1990) and irrelevant phrases that happen to occur in a relevant sentence (Knight and Marcu 2000; Barzilay 2003). Referring expressions in extractive summaries illustrate this, as sentences compiled from different documents might contain too little, too much, or repeated information about the referent.

In a study of how summary revisions could be used to improve cohesion in multi-document summarization (Otterbacher, Radev, and Luo 2002), automatic summaries were manually revised and the revisions classified as pertaining to discourse (34% of all revisions), entities (26%), temporal ordering (22%), and grammar (12%). This study further supports the findings from early research that unclear references in summaries pose serious problems for users (Paice 1990).

2.1.1 Sentence Compression and Fusion. Research in abstractive summarization has largely focused on the problem of compression, developing techniques to edit sentences by removing information that is not salient from extracted sentences. Some approaches use linguistic rules (e.g., Zajic et al. 2007) often combined with corpus-based information (Jing and McKeown 2000), whereas other approaches use statistical compression applied to news (Knight and Marcu 2000; Daumé III and Marcu 2002) and to spoken dialogue (Galley and McKeown 2007). Other researchers addressed the problem of generating new sentences to include in a summary. Information fusion, which uses bottom–up multi-sequence alignment of the parse trees of similar sentences, generates new summary sentences from phrases extracted from different document sentences (Barzilay and McKeown 2005; Filippova and Strube 2008).

2.1.2 Summary Revision. Research in single-document summarization on improving summaries through revision (Mani, Gates, and Bloedorn 1999) is closer to our work. Three types of ad hoc revision rules are defined—*elimination* (removing parentheticals, sentence initial prepositional phrases, and adverbial phrases), *aggregation* (combining constituents from two sentences), and *smoothing*. The smoothing operators cover some reference editing operations. They include substitution of a proper name with a name alias if the name is mentioned earlier, expansion of a pronoun with co-referential proper name in a parenthetical, and replacement of a definite NP with a co-referential indefinite if the definite occurs without a prior indefinite. Mani et al.'s (1999) work differs from ours in that it focuses primarily on subsequent mention (with the exception of pronoun replacement), is meant to work for all entities, not just mentions to people, and does not incorporate distinctions inferred from the input to the summarizer.

Although the rules and the overall approach are based on reasonable intuitions, in practice entity rewrites for summarization do introduce errors, some due to the rewrite rules themselves, others due to problems with co-reference resolution and parsing.

Readers are very sensitive to these errors and prefer extractive summaries to summaries where all references have been edited (Nenkova 2008). Automatic anaphora resolution for all entities mentioned in the input and summary text is also errorful, with about one third of all substitutions in the summary being incorrect (Steinberger et al. 2007). In contrast, when editing references is restricted to references to people alone, as we do in the work presented here, there are fewer edits per summary but the overall result is perceived as better than the original by readers (Nenkova and McKeown 2003).

2.1.3 Reference in Summaries. There has been little investigation of the phenomenon of reference in news summaries. In addition to the revision of subsequent references described in Mani, Gates, and Bloedorn (1999), we are aware of Radev and McKeown (1997), who built a prototype system called PROFILE that extracted references to people from news, merging and recording information about people mentioned in various news articles. The idea behind the system was that the rich profiles collected for people could be used in summaries of later news in order to generate informative descriptions. However, the collection of information about entities from different contexts and different points in time leads to complications in description generation; for example, past news can refer to Bill Clinton as *Clinton, an Arkansas native*, *the democratic presidential candidate Bill Clinton*, *U.S. President Clinton*, or *former president Clinton* and it is not clear which of these descriptions are appropriate to use in a summary of a novel news item. In later work, Radev and McKeown (1998) developed an approach to learn correlations between linguistic indicators and semantic constraints to address such problems, but this line of research has not been pursued further.

Next, we review related work on reference outside the field of summarization.

2.2 Information Status and Generating Referring Expressions

Research on information status distinctions closely relates to work on generating referring expressions. We now overview the two fields and how they interact.

2.2.1 Information Status Distinctions. Information status distinctions depend on two parameters related to the referent's place in the discourse model maintained by the reader: (a) whether it already exists in the hearer's model of the discourse and (b) its degree of salience. The influence of these distinctions on the form of referring expressions has been a focus of past research. For example, centering theory (Grosz, Joshi, and Weinstein 1995) deals predominantly with local salience (local attentional status), and the givenness hierarchy of Prince (1992) focuses on how a referent entered the discourse model (e.g., through a direct mention in the current discourse, through previous knowledge, or through inference), leading to distinctions such as discourse-old, discourse-new, hearer-old, hearer-new, inferable, and containing inferable. Gundel, Hedberg, and Zacharski (1993) attempt to merge salience and givenness in a single hierarchy consisting of six distinctions in cognitive status (in focus, activated, familiar, uniquely identifiable, referential, type-identifiable). In all three theories, familiarity and salience distinctions are shown to be associated with different preferences for syntactic form in the realization of referring expressions.

2.2.2 Generating Referring Expressions (GRE). The most developed sub-area of referring expression generation deals with the problem of generating distinguishing descriptions—descriptions that include enough attributes of the intended referent so that it becomes uniquely identifiable among other entities (Dale 1992). The original

incremental algorithm (Dale and Reiter 1995) assumes that a list of attributes of the discourse entities is readily available and that attributes that rule out the most distractors are added to the referring expression until its interpretation contains only the intended referent. Subsequent work on referring expression generation has (a) expanded the logical framework to allow reference by negation (*the dog that is not black*) and references to multiple entities (*the brown or black dogs*) (van Deemter 2002; Gatt and Van Deemter 2007), (b) explored different search algorithms for finding a minimal description (e.g., Horacek 2003), and (c) offered different representation frameworks such as graph theory (Krahmer, van Erk, and Verleg 2003) or reference domain theory (Denis 2010) as alternatives for representing referring characteristics. This body of research assumes a limited domain where the semantics of attributes and their allowed values can be formalized, though semantic representations and inference mechanisms are getting increasingly sophisticated (e.g., the use of description logic: Areces, Koller, and Striegnitz 2008; Ren, van Deemter, and Pan 2010). In contrast, Siddharthan and Copestake (2004) consider open-domain generation of referring expressions in a regeneration task (text simplification); they take a different approach, approximating the hand-coded domain-knowledge of earlier systems with a measure of relatedness for attribute-values that is derived from WordNet synonym and antonym links.

2.2.3 Recent Trends: Data Collection and Evaluations. There is now increasing awareness that factors other than conciseness are important when planning referring expressions and that considerable variation exists between humans generating referring expressions in similar contexts. Recent evaluation exercises such as the TUNA challenge (Gatt, Belz, and Kow 2008) therefore consider metrics other than length of a reference, such as *humanness* and the time taken by hearers to identify the referent. In a similar vein, Viethen and Dale (2006) examine how similar references produced by well-known algorithms are to human-produced references, and Dale and Viethen (2009) examine differences in human behavior when generating referring expressions. There is also growing collaboration between psycholinguists and computational linguists on the topic of generating referring expressions; for instance, the PRE-CogSci workshop (van Deemter et al. 2010).

Recently, several corpora marked for various information status aspects have been made available. Subsequent studies concerned with predicting givenness status (Nissim 2006; Sridhar et al. 2008), narrow focus (Calhoun 2007; Nenkova and Jurafsky 2007), and rheme and theme distinctions (Postolache, Kruijff-Korbayova, and Kruijff 2005) have not been used for generation or summarization tasks. Current efforts in the language generation community aim at providing a corpus and evaluation task (the GREC challenge) to address just this issue (Belz and Varges 2007; Belz, Kow, and Viethen 2009). The GREC-2.0 corpus, extracted from Wikipedia articles and annotated for the task of referring expression generation for both first and subsequent mentions of the main subject of the article, consists of 2,000 texts in five different domains (cities, countries, rivers, people, and mountains). The more recent GREC-People corpus consists of 1,000 texts in just one domain (people) but references to all people mentioned in a text have been annotated. The GREC challenges require systems to pick the most appropriate reference in context from a list of all references in the document and several defaults, including pronouns, common-noun references, elided reference, and standardized versions of names. By selecting encyclopedic articles about specific referents, this corpus contains large numbers of subsequent references, and in general, the emphasis has been to model the form of subsequent references to named entities in longer texts. Discourse-new vs. discourse-old is the only information status distinction

that participating systems model, with other features derived from lexical and syntactic context; for instance, Greenbacker and McCoy (2009) consider subjecthood, parallelism, recency, and ambiguity.

2.2.4 Applications of Information Status Distinctions to GRE. The main application of theories of information status has been in anaphora resolution. Information status distinctions are not normally used in work on generating referring expressions, with a few notable exceptions.

Krahmer and Theune (2002) show that the relative salience of discourse entities can be taken into account to produce less-informative descriptions (including fewer attributes than those necessary to uniquely identify the referent using a discourse model that does not incorporate salience). In contrast, Jordan and Walker (2005) show that, in task-oriented dialogs, over-specified references (including more attributes than needed to uniquely identify the intended referent) are more likely for certain dialog states and communicative goals. Some participating teams in the GREC challenges use the discourse-new vs. discourse-old distinction as a feature to help select the most likely reference in context. Different interpretations of Centering Theory have also been used to generate pronominal references (McCoy and Strube 1999; Henschel, Cheng, and Poesio 2000).

Our research is substantially different in that we model a much richer set of information status distinctions. Also, our choice of the news genre makes our studies complementary to the GREC challenges, which use Wikipedia articles about people or other named entities. News stories tend to be about events, not people, and the choice of initial references to participants is particularly important to help the reader understand the news. Our research is thus largely focused on the generation of initial references. Due to their short length, summaries do not generally have long co-reference chains and the issue of subsequent reference is of less interest to us. Further, we aim to generate *new* references to people by identifying semantic attributes that are appropriate given the context of the summary. In contrast, the GREC challenges only require the selection of an existing referring expression from a list.

3. Study 1: Discourse-New and Discourse-Old Mentions

Our first study on information status deals with the discourse-new (first mention) versus discourse-old (subsequent mention) distinction. This is the easiest of the three to model, as it is explicitly given in the text. Nevertheless, referring expressions in extractive summaries can be problematic in this respect as sentences compiled from different documents might contain too little, too much, or repeated information about the referent. The first summary reference to a person, for example, may have been the second reference to that person in the input article and thus might not contain enough information to be comprehensible. Conversely, if the second summary reference to a person occurred first in the input article, it may contain more information than needed. In general, in fluent human written text, discourse-new references to entities are longer and more descriptive, whereas discourse-old references are shorter and have a purely referential function. This is not always the case in automatic summaries: Figure 1 shows two extractive summaries. The summaries give a good indication of the problems with reference that can arise in multi-document summaries. In the first summary, references to the former East German Communist leader Erich Honecker are overly repetitive and unnecessary. The extra identification at each mention totals about 15 words, equivalent to the length of an additional informative sentence that could have been included

Summary 1:
Honecker has come under investigation for charges of corruption and living in luxury at the cost of the state. **Former East German leader Erich Honecker** may be moved to a monastery to protect him from a possible lynching by enraged citizens. As protests gathered strength last fall, **Erich Honecker, East Germany's longtime orthodox leader** "lost touch with reality," according to the man who succeeded him as Communist leader only to be ousted later. **Ousted East German leader Erich Honecker**, who is expected to be indicted for high treason, was arrested Monday morning upon release from a hospital and taken to prison.

Summary 2:
Turkey has been trying to form a new government since a coalition government led by **Yilmaz** collapsed last month over allegations that he rigged the sale of a bank. **Ecevit** refused even to consult with the leader of the Virtue Party during his efforts to form a government. **Ecevit** must now try to build a government. **Demirel** consulted Turkey's party leaders immediately after **Ecevit** gave up.

Figure 1
Examples of problematic extractive summaries. The first contains too little information in the first reference and too much repeated detail in subsequent references. The second summary does not provide any information about any of the participants in the story.

instead. In the second summary, the references to Turkish politicians are likely to remain unclear to most readers because of the reduced forms that are realized.

3.1 Corpus Analysis

In order to develop strategies for addressing these issues, we performed a corpus study on the syntactic form of references to people in news text. We used a corpus of news from the test data used in the Document Understanding Conference (DUC) Multi-document Summarization Track (2001–2003), containing 651,000 words and coming from 876 news reports from six different news agencies. The variety of sources is important because working with text from a single source could lead to the learning of paper-specific editorial rules.

The reference characteristics we were interested in were number of pre-modifiers, presence and type of post-modifiers, and the form of name used to refer to people. The corpus was automatically annotated for person name occurrence and co-reference using Nominator (Wacholder, Ravin, and Choi 1997). Syntactic form of references was obtained using Charniak's parser (Charniak 2000). This automatically annotated corpus contains references to 6,240 distinct people.

The distribution of forms for discourse-new and discourse-old references are shown in Table 1. For discourse-old references, computing the probability of a syntactic realization is not as straightforward as for discourse-new references, because the form of the reference is influenced by the form of previous references, among other factors. To capture this relationship, we used the data from discourse-old mentions to form a Markov chain, which captures exactly the probability of transitioning from one form of reference to another. The stationary distributions of the chains for name and pre- and post-modification were computed as an indication of the likelihood of each form in any discourse-old mentions, and are shown in the last column of Table 1.

It is evident from these statistics that, in general, *discourse-new* references should contain the full name and some form of modification, whereas *discourse-old* references should be referred to only by surname. At the same time, there are occasions when the surname only can be used for discourse-new references and, for one in every four references, modifiers are also not necessary. It is thus important to identify referent

Table 1
Likelihood of reference forms for discourse-new and discourse-old references in DUC multi-document track news clusters.

		Discourse-new	Discourse-old
Name Form	Full name	0.97	0.08
	Surname only	0.02	0.87
	Other (e.g., Britney, JLo)	0.01	0.05
Pre-Modification	Any	0.51	0.21
	None	0.49	0.79
Post-Modification	None	0.60	0.89
	Apposition	0.25	0.04
	Relative clause	0.07	0.03
	Other	0.08	0.04
Any Modification (Either Pre- or Post-)	Some Modification	0.76	0.30
	No Modification	0.24	0.70

properties that can help determine when these less-common types of references are felicitous, and we will indeed turn to this question in Section 4.

3.2 Algorithm for Generating Discourse-New and Discourse-Old References to People

To capture the patterns revealed by the corpus analysis, we developed a set of rewrite rules for references to people shown in Algorithm 1. These rules ensure that the discourse-new reference is descriptive and conveys the full name of the person, and that discourse-old references are as brief as possible, using only the surname of the person. For the discourse-new reference, information from all news articles in the summarization input is used. Applying Algorithm 1 to the extractive summaries from Figure 1 produces the rewrite versions of the summaries in Figure 2.

3.3 Evaluation

To evaluate the impact of the rewrite rules on the overall summary, Algorithm 1 was used to rewrite 11 summaries chosen at random from the DUC 2001 and 2002 summaries that contained at least one reference to a person. Four human judges (graduate students at Columbia University, two in computational linguistics and two in other areas) were then given the pairs of the original summary and its rewritten variant without being explicitly told which is which. They were asked to read the summaries and decide if they prefer one text over the other or if they are equal. They were also asked to give free-form comments on what they would change themselves. The distribution of the 44 preferences is as follows: 39 (89%) preferences were for the rewritten form, 4 (9%) were for the original, and there was no preference for the remaining 1 (2%). This preference for the rewritten form is significant (z-test; $p < 0.01$).

In two out of the four cases where the assessor preferred the original version, they commented that the reason for the preference was that the original version exhibited more variation. This observation indicates that the rule for strictly using surname at discourse-old references is too rigid and most probably will need modification in cases where a person is mentioned more often.

Rewrite rules for discourse-new references:

1. IF the person's name is the head of the noun phrase THEN:
   (a) IF any pre-modification is found in the input THEN:
      i. Insert full name and the longest, in number of words, pre-modifier sequence found in the input articles. In case of ties, the discourse-new reference from the article from which the summary sentence is drawn is preferred.
   (b) ELSE IF no pre-modification is found in the input THEN:
      i. Check all discourse-new references in the input to see if any of them includes an apposition modifier.
      ii. Take the longest such modifier and include it in the discourse-new reference NP. Pre-modification is preferred to apposition because they are more frequently used in human-produced texts as shown in the statistics above (in 51% vs. 25% of cases).
2. ELSE IF The name is not the head of the noun phrase it appears in, do not rewrite it.

Rewrite rules for discourse-old references:

1. Use surname only, remove all pre- and post-modifiers.

Algorithm 1: Form of discourse-new vs. discourse-old references.

Summary 1:
Former East German leader Erich Honecker has come under investigation for charges of corruption and living in luxury at the cost of the state. **Honecker** may be moved to a monastery to protect him from a possible lynching by enraged citizens. As protests gathered strength last fall, **Honecker** "lost touch with reality," according to the man who succeeded him as Communist leader only to be ousted later. **Honecker**, who is expected to be indicted for high treason, was arrested Monday morning upon release from a hospital and taken to prison.

Summary 2:
Turkey has been trying to form a new government since a coalition government led by **Prime Minister Mesut Yilmaz** collapsed last month over allegations that he rigged the sale of a bank. **Premier-designate Bulent Ecevit** refused even to consult with the leader of the Virtue Party during his efforts to form a government. **Ecevit** must now try to build a government. **President Suleyman Demirel** consulted Turkey's party leaders immediately after **Ecevit** gave up.

Figure 2
Rewritten versions of extractive summaries from Figure 1.

These results show that even the basic distinction of discourse-new and discourse-old reference in the input and the summaries, the simplest distinction we explore, can help improve summaries. Still, our distributional analysis of syntactic forms in human-generated summaries showed that a full quarter of the discourse-new references contain no modification at all, and half the discourse-new references contained either no modification or just a title or role modifier (e.g., *Mr.* or *President*). There are also differences in the forms of references to people between news reports and human summaries of news (Siddharthan, Nenkova, and McKeown 2004), so not all discourse-new references from the input should be reused directly. Journalistic conventions for many mainstream newspapers dictate that initial mentions to people include a minimum description such as their role or title and affiliation. However, in news summaries written by humans, there are greater space constraints that might warrant shorter references to people. In order to capture reference variation and compression correctly, we need further

distinctions (familiarity and global salience) that will help us determine when it is felicitous to leave out information. We next describe how we learn these distinctions (Section 4) before using them to make generation decisions (Section 5).

4. Study 2: Automatic Prediction of Referent Familiarity and Salience

Deciding how much and what information to include in a reference is influenced by at least two factors: the degree of *familiarity* of the referent (hearer-old vs. hearer-new distinction) and the degree of global *salience* of the referent (is the entity a major or minor character in the news event). These two distinctions have not previously been studied in the context of summarization. There is a trade-off, particularly important for a short summary, between what the speaker wants to convey and how much the listener needs to know. The major/minor distinction plays a role in defining the communication goal (what the summary should be about, which characters are important enough to refer to by name, etc.). The hearer-old/hearer-new distinction can be used to determine whether a description for a character is required from the listener's perspective.

Hearer-old vs. hearer-new. Hearer-new entities in a summary should be described in sufficient detail, whereas hearer-old entities do not require an introductory description. This distinction can have a significant impact on overall length and intelligibility of the produced summaries. Usually, summaries are very short—100 or 200 words for input articles totaling 5,000 words or more. Several people might be involved in a story, which means that if all participants are fully described, little space will be devoted to actual news. In addition, introducing already-familiar entities might distract the reader from the main story (Grice 1975). It can therefore be useful to refer to an entity that can be assumed hearer-old by just a title + surname (e.g., *President Obama*) or by full name only, with no accompanying description (e.g., *Michael Jackson*). Such application of the hearer-old/hearer-new distinction can explain the results from our corpus study of realizations in which we found that 49% of discourse-new references contain only the name or the name and title with no additional descriptive information.

Major vs. minor. Another distinction that human summarizers seem to make is whether a character in a story is a central or a minor one and this distinction can be conveyed by using different forms of referring expressions. It is common to see references in human summaries such as *the dissident's father*. Usually, discourse-new references made solely with a common noun phrase, without the inclusion of the person's name, are employed when the person is not the main focus of a story and is only peripherally related to the main story focus. By automatically inferring if a character is major or minor for the story, we can decide whether to name the character in the summary. Furthermore, many summarization systems use the presence of named entities as a feature for computing the importance of a sentence (Ge, Huang, and Wu 2003; Saggion and Gaizaukas 2004). The ability to identify and use only the major story characters for computing sentence importance can benefit such systems because, in the multi-document summarization track of DUC, only 5% of all people mentioned in the input are also mentioned in the human summaries.

In this section, we report our experiments on building an automatic predictor for the hearer-old/hearer-new and major/minor distinctions. For these experiments, we use data from DUC to approximate the distinctions of interest, without the need for manual annotation. We then validate the performance of the predictor on an independent data set using human judgments as gold standard, achieving high accuracy. These two

distinctions are then used to automatically predict the form and content of discourse-new reference in summaries (Section 5).

4.1 Data Preparation

We use data from the DUC multi-document summarization collection (2001–2004), consisting of 170 pairs of document input sets (10 documents per set) and the corresponding human-written multi-document summaries (two or four per set). Our aim is to identify every person mentioned in a document set (the 10 news reports and the associated human summaries), assign labels for the hearer-old/hearer-new and major/minor distinctions, and generate feature sets for supervised learning. To do this, we first pre-process the data as described next.

4.1.1 Automatic Pre-Processing. Our experiment requires an analysis of every reference to a person in the input documents. For this purpose, we use a variety of tools to perform named entity resolution and co-reference, and to analyze the content of references, including pre-modifiers and post-modification using relative clauses or appositives.

To identify the level of analysis required, we manually inspected the first 500 discourse-new references to people from human summaries in the DUC 2001–2004 data. We found that 71% of pre-modifying words were either title or role words (e.g., *Prime Minister, Physicist,* or *Dr.*) or temporal role modifying adjectives such as *former* or *designate*. Affiliation (country, state, location, or organization names) constituted 22% of pre-modifying words. All other kinds of pre-modifying words, such as *moderate* or *loyal* constituted only 7%. We concluded that there are two attributes that are particularly relevant for generation of discourse-new references of people: role and affiliation.[1]

For named entity recognition and semantic analysis of references, all input documents and summaries were tagged using IDENTIFINDER (Bikel, Schwartz, and Weischedel 1999) to mark up person names, organizations, and locations. We marked up countries and American states using a list obtained from the CIA factsheet.[2] The list consists of 230 country/nationality pairs. To mark up roles, we used a list derived from WordNet (Miller et al. 1993) hyponyms of the *person* synset. Our list has 2,371 entries including multiword expressions such as *chancellor of the exchequer*, *brother in law*, *senior vice president*, and so forth. The list is quite comprehensive and includes roles from the fields of sports, politics, religion, military, business, and many others. We also used WordNet to obtain a list of 58 temporal adjectives. WordNet classifies these as pre- (e.g., *occasional*, *former*, *incoming*) or post-nominal (e.g., *elect*, *designate*, *emeritus*).

In addition, the documents and summaries were annotated with a part-of-speech tagger and simplex noun-phrase chunker (Grover et al. 2000). Also for each named entity, relative clauses, appositional phrases, and copula constructs, as well as pronominal co-reference, were automatically annotated (Siddharthan 2003a, 2003b). The goal was to find, for each person mentioned in the input set, the list of all references to the person in all input documents. For this purpose, all input documents were concatenated and processed with IDENTIFINDER. The IDENTIFINDER output was automatically

1 Our findings are not specific to the news genre or the summarization task; Sekine and Artiles (2009) report that their annotators marked 123 attributes of people mentioned in 156 Web documents. The four most frequent attributes in their collection were: occupation, work, affiliation, and full name; these are the same attributes that we identify.

2 `https://www.cia.gov/library/publications/the-world-factbook/fields/2110.html`, which provides a list of countries and states, abbreviations, and adjectival forms; for example, *United Kingdom/U.K./British/Briton* and *California/Ca./Californian*.

Andrei Sakharov	
Doc 1:	[IR] laureate Andrei D. Sakharov ; [CO] Sakharov ; [CO] Sakharov ; [CO] Sakharov ; [CO] Sakharov ; [PR] his ; [CO] Sakharov ; [PR] his ; [CO] Sakharov ; [RC] who acted as an unofficial Kremlin envoy to the troubled Transcaucasian region last month ; [PR] he ; [PR] He ; [CO] Sakharov ;
Doc 2:	[IR] Andrei Sakharov ; [AP] , 68 , a Nobel Peace Prize winner and a human rights activist , ; [CO] Sakharov ; [IS] a physicist ; [PR] his ; [CO] Sakharov ;

Figure 3
Example information collected for *Andrei Sakharov* from two news reports. "IR" stands for "initial reference", "CO" for noun co-reference, "PR" for pronoun reference, "AP" for apposition, "RC" for relative clause and "IS" for copula.

post-processed to mark up co-referring names and to assign a unique canonical name for each name co-reference chain. For co-reference, a simple rule of matching the surname was used, and the canonical name was the "FirstName LastName" string where the two parts of the name could be identified.[3] Concatenating all documents assures that the same canonical name will be assigned to all named references to the same person.

The tools for pronoun co-reference and clause and apposition identification and attachment (Siddharthan 2002, 2003a) were run separately on each document. Then for each item in the list of canonical names derived from the IDENTIFINDER output, we matched the initial reference in the generic co-reference list for the document with the surname from the canonical name list. The pre-processing steps described previously allow us to collect co-reference information and realization forms (see Figure 3) for each person in each input set, for documents and summaries.

The tools that we used have been evaluated separately when used in a single document setting. In our cross-document matching processes, we could incur more errors, for example, when the co-reference chain in the merged documents is not accurate. On average, out of 27 people per input cluster of documents, 4 people are lost in the matching step for a variety of reasons such as errors in the clause identifier or the co-reference.

4.1.2 Data Labeling. We are interested in acquiring labeled data for familiarity (whether a person is likely to be hearer-old or hearer-new) and global salience (whether a person is a major or minor character in the news story). We now describe how we create labeled data for each of these distinctions.

Hearer-old vs. hearer-new. Entities were automatically labeled as hearer-old or hearer-new by analyzing the syntactic form that human summarizers used for initial references to them. The labeling rests on the assumption that the people who produced the summaries used their own prior knowledge in choosing appropriate references for the summary. Thus, they could refer to people they regarded as familiar to the general public using short forms such as (1) title or role + surname or (2) full name only with no pre- or post-modification. Entities were labeled as hearer-old when the majority of human summarizers for the set referred to them using the forms (1) or (2) (we discuss the validity of this automatic labeling process in Section 4.3.1). The hearer-new/hearer-old distinction is dynamic; when initially unfamiliar characters (like

3 Occasionally, two or more different people with the same surname are discussed in the same set and this algorithm would lead to errors in such cases. We did keep a list of first names associated with the entity, so a more refined matching model could be developed, but this was not the focus of our work.

Saddam Hussein before the first Gulf War) appear in the news over a period of time, they can become hearer-old. Thus the classification of the same person can be different for different document sets dating to different years. From the people mentioned in human summaries, we obtained 118 examples of hearer-old and 140 examples of hearer-new persons—258 examples in total—for supervised machine learning.

Major vs. minor. In order to label an entity as major or minor, we again used the human summaries. Entities that were mentioned *by name* in at least one summary were labeled *major*, and those not mentioned by name in any summary were labeled *minor*. The underlying assumption is that people who are not mentioned in any human summary, or are mentioned without being named, are not central to the story. There were 258 major characters whose names made it to at least one human summary and 3,926 minor characters whose names were absent from all human summaries. The small fraction of major entities in the corpus is not surprising, because many people in news articles express opinions, make statements, or are in some other way indirectly related to the story, without being central to it.

4.2 Machine Learning Experiments

Having created labeled data for classifying people as hearer-new or hearer-old and as major or minor characters, we now proceed to learn these distinctions in a supervised framework. For our experiments, we used the WEKA (Holmes, Donkin, and Witten 1994) machine learning toolkit and obtained the best results for hearer-old/hearer-new using a support vector machine (Sequential Minimal Optimization [SMO] algorithm, with default parameters) and for major/minor, a tree-based classifier (J48, with WEKA parameters: "J48 -U -M 4").

We now discuss what features we used for our two classification tasks (see the list of features in Table 2). Our hypothesis is that features capturing the frequency and syntactic and lexical forms of references are sufficient to infer the desired distinctions.

The frequency features are likely to give a good indication of the global salience of a person in the document set. Pronominalization indicates that an entity was particularly salient at a specific point of the discourse, as has been widely discussed in attentional status and centering literature (Grosz and Sidner 1986; Gordon, Grosz, and Gilliom 1993). Modified noun phrases (with apposition, relative clauses, or pre-modification) can also signal different information status; for instance, we expect post-modification to be more prevalent for characters who are less familiar. For our lexical features, we used two months worth of news articles collected over the Web (and independent of the DUC collection) to collect unigram and bigram lexical models of discourse-new references of people. The names themselves were removed from the discourse-new reference noun phrases and the counts were collected over the pre-modifiers only. One of the lexical features we used is whether a person's description contains any of the 20 most frequent description words from our Web corpus. We reasoned that these frequent descriptors may signal importance; the full list is:

> *president, former, spokesman, sen, dr, chief, coach, attorney, minister, director, gov, rep, leader, secretary, rev, judge, US, general, manager, chairman*

We also used features based on the overall likelihood of a person's description using the bigram model from our Web corpus. These features can help indicate whether a person has a role or affiliation that is important.

Table 2
List of features used for classification.

	Frequency Features		
0,1:	Number of references to the person, including pronouns (total and normalized by feature 2)	2:	Total number of documents containing the person
3:	Proportion of discourse-new references containing full name	4:	Number of times the person was referred to by name after the discourse-new reference
	Syntactic Features		
5,6:	Number of appositives or relative clauses attaching to initial references (total and normalized by feature 2)	7,8:	Number of times apposition was used to describe the person (total and normalized by feature 2)
9,10:	Number of times a relative clause was used to describe the person (total and normalized by feature 2)	11,12:	Number of apposition, relative clause or copula descriptions (total and normalized by feature 2)
13,14:	Number of copula constructions involving the person (total and normalized by feature 2)		
	Lexical Features		
15,16,17:	Probability of an initial reference according to a bigram model (av., max, and min of all initial references)	18:	Number of top 20 high frequency description words (from references to people in a large news corpus) present in initial references

In the experiments reported subsequently, all our features are derived exclusively from the input documents, and we do not derive any features from the summaries. We performed 20-fold cross validation for both classification tasks.

4.2.1 Hearer-Old vs. Hearer-New Results. We present our results for classifying people as hearer-old/hearer-new in Table 3. The 0.54 majority class prediction for the hearer-old/hearer-new classification task is that no-one is known to the reader. Using this prediction in a summarizer would result in excessive detail in referring expressions and a consequent reduction in space available to summarize the news events. The SMO prediction outperformed the baseline accuracy by 22 percentage points (significant at $p = 0.01$, z-test) and is more meaningful for real tasks.

We performed feature selection (using the WEKA CfsSubsetEval attribute evaluator and BestFirst -D 1 -N 5 search method) to identify which are the most important features for this classification task. The important features were: the number of appositions (features 7, 8) and relative clauses (feature 9), number of mentions within the document set (features 0,1), total number of apposition, relative clauses and copula (feature 12), number of high frequency pre-modifiers (feature 18), and the minimum bigram probability (feature 17). Thus, the lexical and syntactic features were more useful than frequency features for determining familiarity.

4.2.2 Major vs. Minor Results. For major/minor classification, the majority class prediction has 94% accuracy (Table 4), but is not useful for a reference generation task as it

Table 3
Cross-validation Accuracy and P/R/F results for hearer-old vs. hearer-new predictions (258 data points). The improvement in accuracy of SMO over the baseline is statistically significant (z-test; $p < 0.01$).

Classifier	Accuracy	Class	Precision	Recall	F
SMO (Only frequency features)	0.64	hearer-new	0.76	0.43	0.55
		hearer-old	0.56	0.84	0.67
SMO (Only lexical features)	0.65	hearer-new	0.64	0.81	0.71
		hearer-old	0.68	0.47	0.55
SMO (Only syntactic features)	0.72	hearer-new	0.82	0.65	0.73
		hearer-old	0.67	0.83	0.74
SMO (frequency+lexical features)	0.66	hearer-new	0.65	0.82	0.72
		hearer-old	0.70	0.48	0.57
SMO (all features)	**0.76**	**hearer-new**	**0.84**	**0.68**	**0.75**
		hearer-old	**0.69**	**0.85**	**0.76**
Majority class prediction	0.54	hearer-new	0.54	1.00	0.70
		hearer-old	0.00	0.00	0.00

Table 4
Cross-validation Accuracy and P/R/F results for major vs. minor predictions (4,184 data points). The improvement in accuracy of J48 over the baseline is statistically significant (z-test; $p < 0.01$).

Classifier	Accuracy	Class	Precision	Recall	F
J48 (Only frequency features)	0.95	major-character	0.81	0.38	0.51
		minor-character	0.96	0.99	0.98
J48 (Only lexical features)	0.94	major-character	0.70	0.16	0.26
		minor-character	0.95	0.99	0.97
J48 (Only syntactic features)	0.95	major-character	0.72	0.36	0.48
		minor-character	0.96	0.99	0.98
J48 (frequency+lexical features)	0.96	major-character	0.69	0.47	0.56
		minor-character	0.96	0.99	0.98
J48 (all features)	**0.96**	**major-character**	**0.70**	**0.53**	**0.60**
		minor-character	**0.97**	**0.99**	**0.98**
Majority class prediction	0.94	major-character	0.00	0.00	0.00
		minor-character	0.94	1.00	0.97

predicts that *no* person should be mentioned by name and all are minor characters. The machine learning approach improves on the baseline accuracy by two percentage points, which is statistically significant (z-test; $p < 0.01$). Due to the skewed nature of our data, precision/recall measures are more useful for analyzing our results. Table 4 shows the performance of the machine learner with different combinations of frequency and lexical and syntactic features. The best results are obtained using all three types of features and it appears that all three aspects are important, yielding an F-measure of 0.60 for the smaller major-character class and and 0.98 for the majority minor-character class.

In a task where ten 400–500 word documents are summarized into 100 words, human summarizers can differ in their interpretations of what is most important to convey. This is a well-established and studied fact in summarization (van Halteren and Teufel 2003). To study how human agreement on the major/minor distinction relates to our automatic prediction results, we further analyzed the 148 persons from DUC

Table 5
J48 Recall results and human agreement for major vs. minor classifications.

Number of summaries containing the person	Number of examples	Number and % recalled by J48
1 out of 4	59	15 (20%)
2 out of 4	35	20 (57%)
3 out of 4	29	23 (79%)
4 out of 4	25	21 (84%)

'03 and DUC '04 sets for which DUC provides four human summaries (there were only two summaries provided for earlier sets). Table 5 presents the distribution of recall taking into account *how many* humans mentioned the person by name in their summary (in our data-labeling, people are labeled as major if *any* summary had a reference to them, see Section 4.1.2). It can be seen that recall is high (0.84) when all four humans consider a character to be major, and falls to 0.2 when only one out of four humans does.

We performed feature selection (using the WEKA CfsSubsetEval attribute evaluator and BestFirst -D 1 -N 5 search method) to identify which are the most important features for the classification task. For the major/minor classification, the important features used by the classifier were the number of documents in which the person was mentioned (feature 2); number of mentions within the document set (features 1, 4); number of relative clauses (feature 9, 10) and copula (feature 13) constructs; total number of apposition, relative clauses, and copula (feature 11); number of high frequency premodifiers (feature 18); and the maximum bigram probability (feature 16).

As for the hearer-old/hearer-new classification, the syntactic forms of references were a significant indicator, suggesting that the centrality of the character was signaled by journalists using specific syntactic constructs in the references. On the other hand, unlike the case of familiarity classification, the frequency of mention within and across documents were also significant features. This is intuitive—a frequently mentioned person is likely to be important to the story.

4.3 Validating the Results on Current News

We tested the classifiers on data different from that provided by DUC, and also tested human consensus on the hearer-new/hearer-old distinction. For these purposes, we downloaded 45 clusters from one day's output from Newsblaster (McKeown et al. 2002). We then automatically compiled the list of people mentioned in the automatic summaries for these clusters. There were 107 unique people that appeared in the automatic summaries and 1,075 people in the input clusters.

4.3.1 Human Agreement on Hearer-Old vs. Hearer-New. A question arises when attempting to infer hearer-new/hearer-old status: Is it meaningful to generalize this across readers, seeing how dependent it is on the world knowledge of individual readers?

To address the question, we gave four American graduate students at Columbia University a list of the names of people in the DUC human summaries (see Section 4.1), and asked them to write down for each person, their country/state/organization affiliation and their role (writer/president/attorney-general, etc.). We considered a

Table 6
Accuracy, precision, and recall for Newsblaster data.

Class	Precision	Recall	F-Measure
Hearer-old	0.88	0.73	0.80
Hearer-new	0.57	0.79	0.66

person hearer-old to a subject if they correctly identified both role and affiliation for that person. For the 258 people in the DUC summaries, the four subjects demonstrated 87% agreement ($\kappa = 0.74$).[4]

Similarly, they were asked to perform the same task for the Newsblaster data, which deals with contemporary news,[5] in contrast with the DUC data that contained news from the late 1980s and early 1990s. On these data, the human agreement was 91% ($\kappa = 0.78$). This is a high enough agreement to suggest that the classification of national and international figures as hearer-old/hearer-new for educated readers is a well-defined task.

4.3.2 Hearer-Old vs. Hearer-New Results on the Newsblaster Data. We measured how well the models learned on DUC data perform with current news labeled using human judgment. For each person who was mentioned in the automatic summaries for the Newsblaster data, we compiled one judgment from the four human subjects using majority vote (an example was labeled as hearer-new if two or more out of the four subjects had marked it as hearer-new; the ties were resolved in favor of hearer-new as it is better to provide an initial description of a person when unsure about the person's status). Then we used these data as *test data*, to test the model trained solely on the DUC data. These results are reported in Table 6. The classifier for hearer-old/hearer-new distinction achieved 75% accuracy on Newsblaster data labeled by humans (significantly better than the majority class (hearer-new) baseline of 60.8%; z-test, $p = 0.02$). This compares well with the reported cross-validation accuracy on DUC data of 76% and indicates that the performance of the classifier is stable and does not vary between the DUC and Newsblaster data. The precision and recall for the Newsblaster data (see Table 6) are also very similar to those for the DUC data.

4.3.3 Major vs. Minor Results on Newsblaster Data. For the Newsblaster data, no human summaries were available, so no direct indication of whether a human summarizer will mention a person by name in a summary was available. In order to evaluate the performance of the classifier, we gave a human annotator (a graduate student at Columbia University) the list of people's names appearing in the machine summaries, together with the input cluster and the machine summary, and asked which of the names on the list would be a suitable keyword for the set. Our aim here was to verify that our classifications of people as major or minor correlate with another indicator of importance—suitability for use as a keyword.

4 κ (kappa) is a measure of inter-annotator agreement over and above what might be expected by pure chance (see Carletta [1996] for discussion of its use in NLP). $\kappa = 1$ if there is perfect agreement between annotators, $\kappa = 0$ if the annotators agree only as much as you would expect by chance, $\kappa < 0$ if the annotators agree less than predicted by chance.

5 The human judgments were made within a week of the publication of the news stories in the Newsblaster clusters.

Out of the 107 names on the list, the annotator chose 42 as suitable for descriptive keyword for the set. The major/minor classifier was run on these 107 examples; only 40 were predicted to be major characters. Of the 67 test cases that were predicted by the classifier to be minor characters, 12 (18%) were marked by the annotator as acceptable keywords. In comparison, of the 40 characters that were predicted to be major characters by the classifier, 30 (75%) were marked as possible keywords. If the keyword selections of the annotator are taken as ground truth, the automatic predictions have precision and recall of 0.75 and 0.71, respectively, for the *major class*.

5. Using Automatically Inferred Distinctions for Generation Decisions

Having trained models to predict whether a person is a major or a minor character, and whether the person is likely to be hearer-old or hearer-new to the intended audience, we can now make informed decisions on how to generate initial references to people in summaries. In this section, we demonstrate the predictive power of the distinctions we have acquired, by showing how we can determine when to include the name attribute (Section 5.1); post-modification such as apposition or relative clauses (Section 5.2); and pre-modification using specific semantic attributes such as affiliation, role, and temporal modifiers (Section 5.3). Then, in Section 6, we present and evaluate our full algorithm for generating referring expressions to people in multi-document summaries.

5.1 Decision 1: Including the Name Attribute

According to our theory, only major characters should be named in a summary. In addition to using up words, naming minor characters can mark them as being important and distract the reader from the main story by introducing Gricean implicatures (Grice 1975).

In our data, there were 258 people mentioned by name in at least one human summary. In addition to these, there were 103 people who were mentioned in at least one human summary using only a common noun reference (these were identified by hand, as common noun co-reference cannot be performed reliably enough by automatic means). This means that 29% of people mentioned in human summaries are not actually named. Examples of such references include *an off duty black policeman*, *a Nigerian born Roman catholic priest*, and *Kuwait's US ambassador*.

Our WEKA machine learner for the major/minor distinction achieved a testing accuracy of 74% on these 103 examples. In other words, we can reproduce human judgment on which people to refer to by name in three quarters of cases. This is a very encouraging result given the novelty of the task.

As mentioned before, different human summarizers can sometimes make different decisions on the form of reference to use. Out of the 103 examples of people with an unnamed reference in at least one human summary, there were 63 people who were not mentioned by name in any summary. WEKA correctly labeled 58 (92%) as minor characters. Out of the 40 cases where some summarizers used named reference and others used common noun reference, 22 of these 40 (55%) were labeled as minor characters. As before, we observe that when human summarizers generate references of the same form (reflecting consensus on conveying the perceived importance of the character), the machine predictions are very accurate.

5.2 Decision 2: Elaborating Using Post-Modification

One aspect of reference generation that is informed by the hearer-old/hearer-new status is the use of apposition or relative clauses for elaboration. It has been observed (Siddharthan, Nenkova, and McKeown 2004) that, on average, these constructs occur 2.3 times *less* frequently in human summaries than in machine summaries. Post-modification tends to be more lengthy than pre-modification, and by predicting when this is not required, we can achieve large reductions in length.

To determine when an appositive or relative clause can be used to modify a reference, we considered the 151 examples out of 258 where there was at least one relative clause or apposition describing the person in the input. We labeled an example as positive if *at least* one human summary contained an apposition or relative clause for that person and negative otherwise. There were 66 positive and 85 negative examples. This data is informative because although for the majority of examples (56%) all the human summarizers agreed not to use post-modification, there were very few examples (under 5%) where all the humans agreed to post-modify. This reflects the high cost in word count for using these forms of modification. Intuitively, it would appear that for around half the cases (56%), it should be obvious that no post-modification is required, but for the other half, opinions can vary.

We report that *none* of the hearer-old persons (as classified by the SMO algorithm) were post-modified. Our predictions cleanly partition the examples into those where post-modification is not required, and those where it might be. Because we could not think of any simple rule that handled the remaining examples, we added the testing predictions of hearer-old/hearer-new and major/minor as features to the list in Table 2 and tried to learn this task using the tree-based learner J48. We report a testing accuracy of 71.5%, which is significantly higher than both the 56% for the majority class baseline (z-test; $p < 0.01$) and 62.5% for a baseline using the original feature set without the two information status features (z-test; $p < 0.05$).

There were only three useful features—the predicted hearer-new/hearer-old status, the number of high frequency pre-modifiers for that person in the input (feature 18 in Table 2), and the average number of post-modified initial references in the input documents (feature 12).

5.3 Decision 3: Including Pre-Modifying Attributes

As mentioned in Section 4.1.1, our analysis of pre-modification in initial references to people in DUC human summaries showed that 71% of pre-modifying words were either title or role words or temporal role modifying adjectives. Affiliations constituted 22% of pre-modifying words and all other pre-modifying words, such as *moderate* or *loyal* constituted only 7%. We therefore only consider the inclusion of roles, temporal modifiers, and affiliations in this section.

5.3.1 Including Role and Temporal Modification Attributes. The DUC human summarizers tended to follow journalistic conventions regarding the inclusion of a title or role in initial references to people. Indeed a simple rule—to always include the role/title in initial references—reproduced the choices made by the human summarizers in 79% of cases. A manual analysis of cases where human summarizers omitted title/role words revealed some insights. There were a small number of historical figures (e.g., *Galileo* and *Napoleon*) and people from the entertainment industry (e.g., *Robert Redford* and *Yoko Ono*) who were always referred to only by name. Otherwise, the main factor appears to be

notoriety. People who were almost never referred to with a title or role include *Moammar Gadhafi, Osama bin Laden, Fidel Castro, Yasser Arafat*, and *Boris Yeltsin*. Others who were referred to both with and without a title/role by different human summarizers include *George Bush, Bill Clinton, Margaret Thatcher, Michael Gorbachev*, and *Slobodan Milosevic*. As we have no insights as to how to model notoriety, we did not try to improve on this "always include" baseline, but we can nonetheless suggest that for greater compression, the role or title can be omitted for hearer-old persons; for example, generating *Margaret Thatcher* instead of *Former Prime Minister Margaret Thatcher*.

5.3.2 Including Affiliation Attributes. We now describe a procedure that uses hearer and discourse information to decide when to provide an affiliation in the initial reference to a person. This issue is ubiquitous in summarizing news; for example, the reference generator might need to decide between *White House Press Secretary James Brady* and *Press Secretary James Brady*, between *Soviet President Gorbachev* and *President Gorbachev*, or between *Indiana Senator Dan Quayle* and *Senator Dan Quayle*.

1. IF:
   (a) the person is classified as hearer-old OR
   (b) the person's organization (country/ state/ affiliation) has been already mentioned AND is the most salient organization in the discourse at the point where the reference needs to be generated

   THEN the affiliation of a person can be omitted in the discourse-new reference.

Algorithm 2: Omitting the affiliation in a discourse-new reference.

Based on our intuitions about discourse salience and information status, we initially postulated the decision procedure in Algorithm 2. We described how we make the hearer-new/hearer-old judgment in Section 4.2. We used a salience-list (S-List) (Strube 1998) to determine the salience of organizations. This is a shallow attentional-state model and works as follows:

1. Within a sentence, entities are added to the salience-list from left to right.
2. Within the discourse, sentences are considered from right to left.

In other words, entities in more recent sentences are more salient than those in previous ones, and within a sentence, earlier references are more salient than later ones.

Results. To make the evaluation meaningful, we only considered examples where there was an affiliation mentioned for the person in the input documents, ruling out the trivial cases where there was no choice to be made (i.e., an affiliation could never be included). There were 272 initial references to 182 persons in the human summaries that met this criterion (note that there were multiple human summaries for each document set).

We used 139 of these 272 examples (from DUC '01, '02, and '03) as training data to check and possibly refine our rule. For each of these 139 initial references to people, we:

1. Obtained from the source news reports the test-set prediction from WEKA on whether that person was hearer-new or hearer-old.

2. Formed the S-List for affiliations in that human summary at the point of reference.[6]
3. Used the decision procedure in Algorithm 2 to decide whether or not to include the affiliation in the reference.

The evaluation consisted of matching our predictions with the observed references in the human summaries. Our decision procedure made the correct decision in 71% of the instances and successfully modeled variations in the initial references used by different human summarizers for the same document set:

1. **Brazilian President Fernando Henrique Cardoso** was re-elected in the...
   [*hearer-new* and Brazil not in context]
2. Brazil's economic woes dominated the political scene as **President Cardoso**...
   [*hearer-new* and Brazil most salient country in context]

It also modeled variation in initial references to the same person across summaries of different document sets:

1. It appeared that **Iraq's President Saddam Hussein** was determined to solve his country's financial problems and territorial ambitions...
   [*hearer-new* for this document set and Iraq not in context]
2. ...A United States aircraft battle group moved into the Arabian Sea. **Saddam Hussein** warned the Iraqi populace that United States might attack...
   [*hearer-old* for this document set]

An error analysis showed that in most of these instances the rule predicted no affiliation in instances where the human summarizer had included it. In many cases, the person was first mentioned in a context where a different organization/state or country was more salient than their own. When we modified condition (1) of our decision rule (Algorithm 2) to obtain Algorithm 3, the accuracy increased to 78%. The improved performance of our second decision procedure suggests that affiliation is sometimes included in references to even hearer-old persons in order to aid the hearer in immediately recollecting the referent. Both algorithms make errors on such cases, however, and there appears to be some variability in how human summarizers make their decisions in these contexts.

1. IF:
   (a) the person is hearer-old, *and no country/state/org is more salient than their own* OR
   (b) the person's organization (country/ state/ affiliation) has been already mentioned AND is the most salient organization in the discourse at the point where the reference needs to be generated

   THEN the affiliation of a person can be omitted in the discourse-new reference.

Algorithm 3: Omitting the affiliation in a discourse-new reference (version 2).

Having convinced ourselves of the validity of these rules, we applied them to the 133 examples in the unseen test data. The results are shown in Table 7. A total of 85%

6 `http://www.cia.gov/cia/publications/factbook` provides a list of countries and states, abbreviations and adjectival forms; and the named entity recognition tool IDENTIFINDER marks up organizations. The output was manually cleaned to remove errors in named entity detection.

Table 7
Test set results for the decision procedure to include affiliation in initial references. Both rules are significantly better than all the baselines (z-test; $p < 0.05$).

Algorithm	Accuracy
Never-Include Baseline	0.56
Information-Status Baseline	0.58
Salience Baseline	0.65
Salience+Information Status (Algorithm 2)	0.79
Salience+Information Status (Algorithm 3)	0.75

of the observed human references were modeled correctly by either Algorithm 2 or Algorithm 3, demonstrating that the hearer-old/hearer-new distinction is relevant to reference generation. The remaining errors were largely due to misclassifications of people as hearer-new by SMO, thus leading our rule to include affiliation when not required. We compared the rules' prediction accuracy to that of three baselines (see Table 7):

1. **Never-Include**: This is the majority class baseline that says that affiliation is always omitted.
2. **Information-Status**: Always include if hearer-new, never include if hearer-old (using testing predictions from automatic classification of information status).
3. **Salience**: Include affiliation unless that affiliation is already the most salient at the point of reference.

Algorithm 2 performs significantly better than all baselines (z-test; $p < 0.01$), whereas Algorithm 3 performs significantly better than the first two baselines (z-test; $p < 0.01$) and for the third (z-test; $p < 0.05$).

6. An Algorithm for Generating References to People in Summaries

Having shown that important information status distinctions can be acquired automatically (Section 4) and that these distinctions predict aspects of the content and form of references (Section 5), we now update Algorithm 1 from Section 3 to obtain Algorithm 4, our full algorithm for generating references to people in summaries that takes into account the discourse-new vs. discourse-old, hearer-new vs. hearer-old, and major vs. minor distinctions. Table 8 summarizes the accuracy of this algorithm in predicting human generation decisions, as reported in Sections 3 and 5. Next, we report an evaluation of the extent to which reformulating references to people impacts on the quality of news summaries.

6.1 Evaluation of Summaries Rewritten According to Algorithm 4

We evaluated our algorithm using 14 news clusters from the Google News world news section.[7] We selected these clusters in the order they were presented on the Google News site; we excluded three clusters that we deemed too similar to already selected

7 http://news.google.com/news/section?&topic=w.

Rewrite rules for discourse-new references:

1. IF the person's name is the head of the noun phrase THEN:
   (a) IF Minor Character THEN:
      i. EXCLUDE name from reference and only INCLUDE role, temporal modification, and affiliation
   (b) ELSE IF Major Character AND Hearer-old THEN:
      i. INCLUDE name
      ii. INCLUDE role and any temporal modifier, to follow journalistic conventions
      iii. EXCLUDE other modifiers including affiliation
      iv. EXCLUDE any post-modification such as apposition or relative clauses
   (c) ELSE IF Major Character AND Hearer-new THEN:
      i. INCLUDE name
      ii. INCLUDE role and any temporal modifier, to follow journalistic conventions
      iii. IF the person's affiliation has already been mentioned AND is the most salient organization in the discourse at the point where the reference needs to be generated THEN EXCLUDE affiliation ELSE INCLUDE Affiliation
      iv. Use machine learner described in Section 5.2 to decide whether to include post-modification
2. ELSE IF The name is not the head of the noun phrase it appears in, THEN it is not rewritten

Rewrite rules for discourse-old references:

1. Use surname only, EXCLUDE all pre-modifiers and post-modifiers

Algorithm 4: Generating references to people in news summaries.

clusters, however. We used the first 10 articles in each cluster to create our 14 document sets. We then used the freely available extractive summarizer MEAD (Radev et al. 2004) to generate 200 word summaries for each document set. These extractive summaries were automatically rewritten according to Algorithm 4, as described subsequently. Our evaluation compares the extractive and rewritten summaries.

Table 8
Summary of the accuracy of Algorithm 4 for specific regeneration decisions.

Generation Decision	Section	Prediction Accuracy
	Discourse-new references	
Include Name	Section 5.1	.74 (rising to .92 when there is unanimity among human summarizers)
Include Role & temporal mods	Section 5.3.1	.79
Include Affiliation	Section 5.3.2	.75 to .79 (depending on rule)
Include Post-Modification	Section 5.2	.72 (rising to 1.00 when there is unanimity among human summarizers)
	Discourse-old references	
Include Only Surname	Section 3	.70

Implementation of Algorithm 4. Our reference rewrite module operates on parse trees obtained using the Stanford Parser (Klein and Manning 2003). For each person automatically identified using the techniques described in Section 4.1.1, we matched every mention of their surname in the parse trees of MEAD summary sentences. We then replaced the enclosing NP (includes all pre- and post-modifying constituents) with a new NP generated using Algorithm 4. The regenerated summary was produced automatically, without any manual correction of parses, semantic analyses, or information status classifications. We now enumerate implementation details not covered in Section 4.1.1:

1. Although our algorithm determines when to include role and affiliation attributes, it doesn't inform us as to which ones to include; for instance, a politician might have a constituency, party, or nationality affiliation. Our implementation selects the most frequently mentioned role (e.g., *prime minister*) in references to that person in the document set, and then selects the affiliation associated with that role (e.g., *British*).
2. Another situation that arises is when our algorithm prescribes the inclusion of affiliation (or role), but our semantic tagger does not find any affiliation (or role) for that person. In these cases, we check whether there exists any relative clause or apposition, as post-modification is often used to introduce people without using role or affiliation attributes. If no post-modification is available, we include the longest initial reference to the person from the input documents.
3. Different conventions exist regarding which name of a person to use for reference. For instance, in China it is traditional to use the first name (e.g., Chinese Vice Premier Li Keqiang is commonly referred to in news articles as *Li*). We do not claim to model such naming conventions; rather we use the co-reference chains in our analysis to pick the most frequent name used in co-reference (see [CO] tags in Figure 3).

In total, there were 61 references to people modified in the 14 summaries. Of those, 34 involved shortening subsequent references. For initial references, there were 6 instances of removing affiliations, 4 of adding affiliations, and 10 of adding roles. There were also 6 instances of post-modifications added to initial references and 1 removed.

Experimental design. The evaluation was carried out over the Internet using an interface that, on each slide, showed the two summaries (before and after reference rewriting) side by side with the left one labeled A and the right one B. Underneath the summaries, there were three multiple choice questions and one free text question. A screen-shot is shown in Figure 4. The order of presentation of summaries was controlled for (i.e., for each summary pair, equal numbers of participants saw the regenerated summary on the left and on the right, and for each participant, summary order on each slide was pseudo-randomized). In order to prevent evaluator fatigue, we split our 14 summary pairs into two sets of 7 pairs, with each participant evaluating only one set. Participants were provided with the following instructions:

> To help our research into summarizing news stories, please compare the following pairs of summaries, 7 in total. On each slide, the two summaries are quite similar, but we would like you to tell us which you prefer, and which is more informative and coherent. In addition, we would appreciate any subjective assessment of the summaries, particularly with regard to the amount of information provided about participants in the news story.

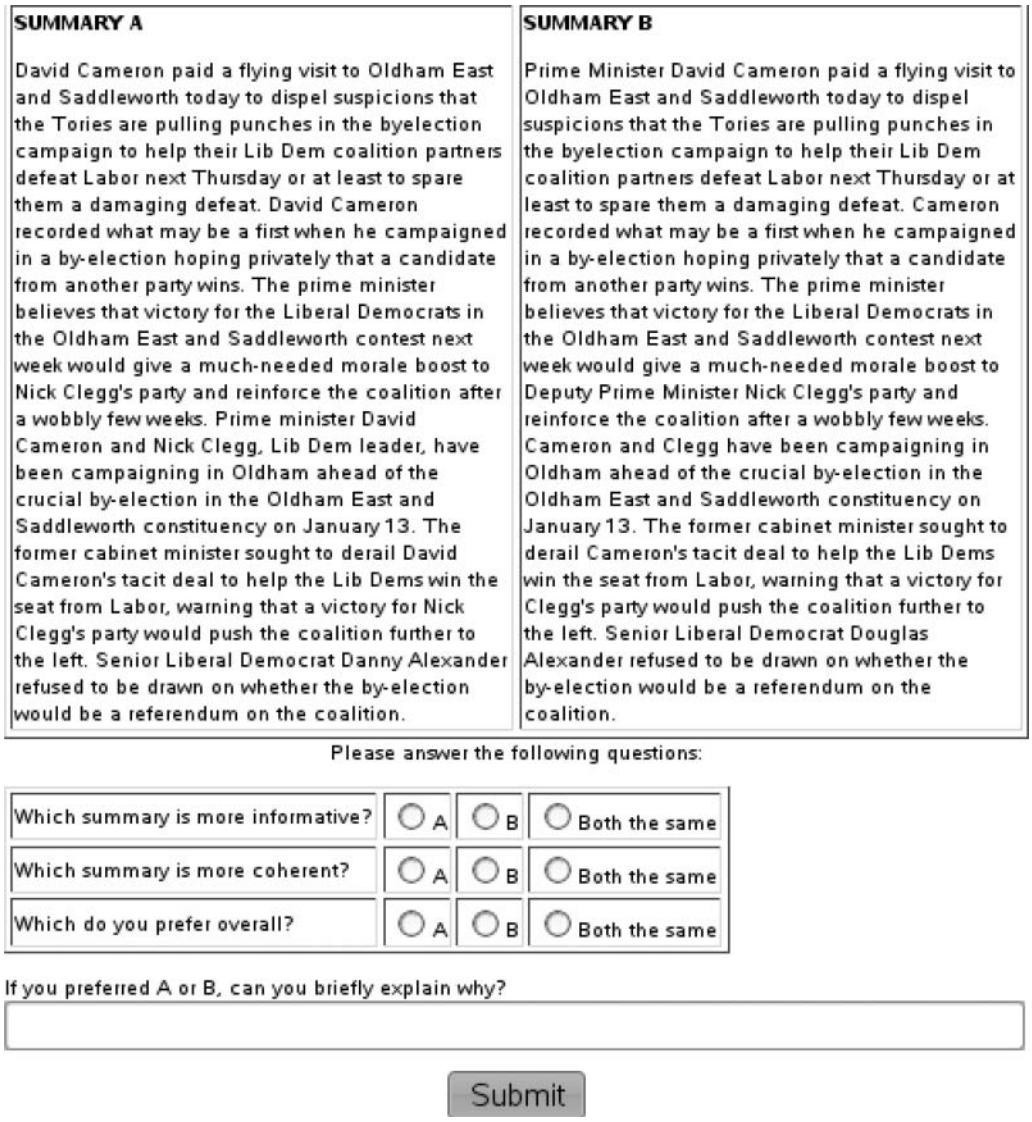

Figure 4
Screen shot of evaluation interface.

Results. Twenty participants (undergraduate and postgraduate students and research fellows at Columbia University, the University of Pennsylvania, and the University of Aberdeen) completed the evaluation. The results are summarized in Table 9. Our evaluation shows that reference rewriting makes sentences more coherent (significant at $p < 0.01$, z-test, sample size = 140), and that this is preferred by participants (significant at $p < 0.01$, z-test, sample size = 140). The loss of informativeness (significant at $p < 0.01$, z-test, sample size = 140) through our rewrites is not unexpected; in general we are removing information from references, and we only rarely add information to a summary. This is evident from the summary lengths; the rewritten summary is shorter for 11 out of 14 document sets and the average word lengths of the extractive and rewritten summaries are 189 and 178, respectively. Indeed, we found a strong correlation between

Table 9
Results for the evaluation for rewritten summaries.

(a) Number of times any participant selected one summary type over the other (140 comparisons):

	More informative	More coherent	More preferred
Extractive	46	22	37
Rewritten	23	79	69
No difference	71	39	34

(b) Number of document sets for which participants selected one summary type more often than the other (14 document sets):

	More informative	More coherent	More preferred
Extractive	9	2	2
Rewritten	4	12	10
Equal	1	0	2

differences in informativeness and differences in summary lengths (Spearman's rho = .79; significant at $p < 0.001$). We did not find any similar correlation between differences in summary lengths and either coherence (Spearman's rho = .05, $p = 0.86$) or overall preference (Spearman's rho = .25, $p = 0.39$).

We also performed repeated measures ANOVAs on coherence, informativeness, and overall preference, with summary-type (Extractive or Rewritten) as a within-participant variable. For this purpose, we converted our categorical user responses ("A", "B", or "Both the same") into numerical ratings for each summary-type condition as follows. We first mapped the choice of "A" or "B" to the summary-type and then:

- If the user selected Extractive, we used a rating of 1 for Extractive and -1 for Rewritten.
- If the user selected Rewritten, we used a rating of 1 for Rewritten and -1 for Extractive.
- If the user selected "Both the same", we used a rating of 0 for both conditions.

Then, treating participants (F1) and news clusters (items; F2) as random factors, we found a main effect of summary-type on Coherence ($F1(1, 19) = 34.20$, $p < 0.001$; $F2(1, 13) = 3.91$, $p = 0.001$). This analysis confirms that the improvement in coherence is significant even when the variation between participants and items (news clusters) is taken into account. We found a smaller by-participant effect of summary-type on Informativeness, and no by-item effect ($F1(1, 19) = 6.12$, $p = 0.023$; $F2(1, 13) = 2.24$, $p = 0.157$). For overall preference, we found a main by-participant effect and a smaller by-item effect of summary-type ($F1(1, 19) = 10.49$, $p = 0.004$; $F2(1, 13) = 3.23$, $p = 0.09$).

6.2 Discussion of Limitations and Scope of Algorithm 4

The evaluation described in the previous section provides us with some feedback about the limitations of our approach, regarding the factors we consider in our algorithm as

well as the automatic analysis we require to fully implement it. We had requested participants to explain their decisions, and this free text feedback proved very informative. We summarize what we learned:

- Sentence length is an important factor. Introducing post-modification into a sentence that was already long, or removing it from a sentence that was already short, resulted in a dispreference for the rewritten summary. Note that sentence length is not a feature we had given any consideration to in our algorithm.
- The lack of common noun co-reference in our automatic analysis cost us. In one summary, there were common noun references to *the widely recognized winner of Ivory Coast's election* and *the internationally recognized winner of Ivory Coast's presidential election*. Because our analysis was unable to co-refer these to the named reference *Alassane Ouattara*, the reference rewrite module actually introduced more redundancy into the summary.
- Our strategy of selecting the most frequent role from the input was not optimal. In more than one instance, this strategy selected the role *leader*, although the original summary had a more informative role such as *prime minister*. This was commented on and penalized by multiple participants for informativeness and often for overall preference as well. Indeed, we found our participants to be very sensitive to the choice of role modifier, consistently penalizing summaries that omit the more specific role (there were 17 comments to this effect, in contrast there were only 3 comments complaining about lack of affiliation).
- Mistakes introduced during automated analysis cost us. When *the Obama administration slapped wide-ranging sanctions* got rewritten as *Obama slapped wide-ranging sanctions* due to incorrect NP matching, the latter was deemed biased and misleading. There was one particular mistake in role identification that got penalized by all participants when the algorithm generated *spin doctor David Cameron*. The phrase in the input document that got misanalyzed was *Andy Coulson, David Cameron's spin doctor*.
- Participants often disagree. Two different participants provided the following comments: *...repetition of ref expressions clarify some ambiguities* and *...introduced differently in each sentence, and that made it harder to see that it was the same person*.

In addition to these limitations that are mostly concerned with implementation issues, we need to reiterate that our claims are genre-specific. We have studied the nature of references to people in the newswire and news summary genres. This is in contrast to many other studies on reference that make use of experimental data from human reference tasks (e.g., Gatt, Belz, and Kow 2008), corpus data from dialog (e.g., Gupta and Stent 2005), or encyclopedic texts (e.g., Belz, Kow, and Viethen 2009). There are big differences in how people are referred to in different genres, arising from both linguistic conventions and the nature of the information in the genre. For instance, encyclopedic or biographical texts about a person contain long co-reference chains, and information about the person is provided throughout the article. Thus, subsequent references in such genres are not straightforward, and research arising from the GREC challenges has thus justifiably focused on modeling subsequent references.

The nature of references to people in the news genre is quite different. As news articles tend to be about events rather than people, there tends to be less information provided about the people involved in the story, and these descriptions are almost always provided in the first reference to the person (see Study 1, Section 3). Thus in the news genre, unlike encyclopedic texts, the task of content selection for subsequent references to people is rather uninteresting. This is even more so for the news summary genre, where co-reference chains are typically short, resulting in few subsequent references to anybody. We find that for the news summary genre, the form and content of initial reference is critical, and our studies have thus focused on this. Specifically, due to the nature of our genre, we do not attempt to model anaphoric phenomena such as pronouns and common noun co-reference. The studies reported in this article should thus be seen as complementary to efforts such as GREC, and we believe that such studies of reference in different genres are important to get a better understanding of the phenomenon.

7. Conclusions

Our research both provides a characterization of references to people in the news genre through empirical analysis and uses that characterization to develop a model for generating references to people in news summaries that is based on automatically inferred information status distinctions. Because summarization takes its content from input full text articles, one contribution of our work is the development of a statistical model for inferring such distinctions from news reports, without requiring manual annotation. We have shown how this model can then be used to generate appropriate references, including semantic content selection (inclusion of name, role, and affiliation attributes) and realization choices (use of pre- or post-modification).

References to people have very different properties from other kinds of referring expressions (e.g., common noun references to objects). Research on the generation of referring expressions from semantic representations is based on the notion of selecting attributes that distinguish the object from others; in contrast, discourse-new references to people often contain attributes in addition to the name, and yet the name alone would be a distinguishing attribute. In this article, we have conducted corpus-based studies to provide answers to how and when attributes are used in references to people in the news genre. Our study characterizes the differences in discourse-new and discourse-old references, identifies the attributes typically used in discourse-new references to people, and provides evidence that information status distinctions (global salience and familiarity) can determine when people are named in summaries and when additional attributes such as role and affiliation are needed.

These information status distinctions are important when generating summaries of news, as they help determine both what to say and how to say it. However, using these distinctions for summarization requires inferring information from unrestricted news. We have shown that the hearer-old/hearer-new and major/minor distinctions can be inferred reliably using features derived from the lexical and syntactic forms and frequencies of references in the news reports. These acquired distinctions are useful for determining which characters to name in summaries, which characters to further describe or elaborate on, and the forms that any descriptions should take.

Finally, we have reported an evaluation of the effect of reference rewriting on extractive summaries, demonstrating that our rewrites improve coherence and are generally preferred by readers.

References

Areces, Carlos, Alexander Koller, and Kristina Striegnitz. 2008. Referring expressions as formulas of description logic. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 42–49, Salt Fork, OH.

Barzilay, Regina. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

Barzilay, Regina and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Belz, Anja, Eric Kow, and Jette Viethen. 2009. The GREC named entity generation challenge 2009: overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 88–98, Suntec, Singapore.

Belz, Anja and Sebastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)*, pages 9–16, Schloss Dagstuhl, Germany.

Bikel, Daniel, Richard Schwartz, and Ralph Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231.

Calhoun, Sasha. 2007. Predicting focus through prominence structure. In *Proceedings of Interspeech'07*, pages 622–625, Antwerp, Belgium.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Conference of North American Chapter of the Association for Computational Linguistics*, pages 132–139, Seattle, WA.

Dale, Robert. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.

Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Dale, Robert and Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 58–65, Athens, Greece.

Daumé III, Hal, Arda Echihabi, Daniel Marcu, Dragos Munteanu, and Radu Soricut. 2002. GLEANS: A generator of logical extracts and abstracts for nice summaries. In *Proceedings of the Second Document Understanding Conference (DUC 2002)*, pages 9–14, Philadelphia, PA.

Daumé III, Hal and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL – 2002)*, pages 449–456, Philadelphia, PA.

Denis, Alexandre. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference-INLG*, pages 27–36, Dublin, Ireland.

Fellbaum, Christine. 1998. WordNet. An electronic lexical database. Cambridge, MA: MIT Press.

Filippova, Katja and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 177–185.

Galley, Michel and Kathleen McKeown. 2007. Lexicalized Markov grammars for sentence compression. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 180–187.

Gatt, Albert, Anja Belz, and Eric Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 198–206.

Gatt, Albert and Kees Van Deemter. 2007. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, 16(4):423–443.

Ge, Jiayin, Xuanjing Huang, and Lide Wu. 2003. Approaches to event-focused summarization based on named entities and query words. In *Document Understanding Conference (DUC'03)*.

Gordon, Peter, Barbara Grosz, and Laura Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Greenbacker, Charles F. and Kathleen F. McCoy. 2009. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on the Production of Referring Expressions: Bridging the Gap*

between Computational and Empirical Approaches to Reference (PRE-CogSci 2009).
Grice, Paul. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics*, volume 3. Academic Press, New York, pages 43–58.
Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.
Grosz, Barbara and Candice Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204.
Grover, Claire, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT: A flexible tokenization toolkit. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 1147–1154, Athens, Greece.
Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.
Gupta, Surabhi and Amanda J. Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Corpora in NLG*, pages 1–6, Birmingham.
Henschel, Renate, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th Conference on Computational Linguistics (COLING'2000)*, pages 306–312, Saarbrücken, Germany.
Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. 1994. Weka: A machine learning workbench. In *Proceedings of the 2nd Australian and New Zealand Conference on Intelligent Information Systems*, pages 357–361, Brisbane, Australia.
Horacek, Helmut. 2003. A best-first search algorithm for generating referring expression. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 103–106, Budapest, Hungary.
Jing, Hongyan and Kathleen McKeown. 2000. Cut and paste based text summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, pages 178–185, Seattle, WA.
Jordan, Pamela and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.
Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.
Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization—step one: Sentence compression. In *Proceedings of The American Association for Artificial Intelligence Conference (AAAI-2000)*, pages 703–710, Austin, TX.
Krahmer, Emiel and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, pages 223–264.
Krahmer, Emiel, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.
Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.
Mani, Inderjeet, Barbara Gates, and Eric Bloedorn. 1999. Improving summaries by revising them. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 558–565, College Park, MD.
McCoy, Kathleen and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of ACL'99 Workshop on the Relationship Between Discourse/Dialogue Structure and Reference*, pages 63–72, College Park, MD.
McKeown, Kathleen, Regina Barzilay, David Evans, Vasleios Hatzivassiloglou, Judith Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the 2nd Human Language Technologies Conference HLT-02*, pages 280–285, San Diego, CA.
Nenkova, Ani. 2008. Entity-driven rewrite for multi-document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP08)*, pages 118–125, Hyderabad, India.

Nenkova, Ani and Dan Jurafsky. 2007. Automatic detection of contrastive elements in spontaneous speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 201–206, Kyoto, Japan.

Nenkova, Ani and Kathleen McKeown. 2003. References to named entities: a corpus study. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 70–72, Edmonton, Canada.

Nenkova, Ani, Advaith Siddharthan, and Kathleen McKeown. 2005. Automatically learning cognitive status for multi-document summarization of newswire. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 241–248, Vancouver, British Columbia, Canada.

Nissim, Malvina. 2006. Learning information status of discourse entities. In *Proceedings of the Conference on Emprical Methods in Natural Language Processing (EMNLP)*, pages 94–102, Sydney, Australia.

Otterbacher, Jahna C., Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 27–36, Philadelphia, PA.

Paice, Chris D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing Management*, 26(1):171–186.

Postolache, Oana, Ivana Kruijff-Korbayova, and Geert-Jan Kruijff. 2005. Data-driven approaches for information structure identification. In *Proceedings of HLT/EMNLP*, pages 9–16, Vancouver, Canada.

Prince, Ellen. 1992. The ZPG letter: subject, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*. John Benjamins, Amsterdam, The Netherlands, pages 295–325.

Radev, Dragomir, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD—A platform for multidocument multilingual text summarization. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC'2004)*, pages 699–702, Lisbon, Portugal.

Radev, Dragomir and Kathleen McKeown. 1997. Building a generation knowledge source using internet-accessible newswire. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 221–228, Washington, DC.

Radev, Dragomir and Kathleen McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Ren, Yuan, Kees van Deemter, and Jeff Pan. 2010. Charting the potential of description logic for the generation of referring expressions. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 115–124, Dublin, Ireland.

Saggion, Horacio and Rob Gaizaukas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Document Understanding Conference (DUC04)*.

Sekine, Satoshi and Javier Artiles. 2009. Weps2 attribute extraction task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, Madrid, Spain.

Siddharthan, Advaith. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, PA.

Siddharthan, Advaith. 2003a. Resolving pronouns robustly: Plumbing the depths of shallowness. In *Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 7–14, Budapest, Hungary.

Siddharthan, Advaith. 2003b. *Syntactic Simplification and Text Cohesion*. Ph.D. thesis, University of Cambridge, UK.

Siddharthan, Advaith and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42th Meeting of the Association for Computational Linguistics Annual Conference (ACL 2004)*, pages 407–414, Barcelona, Spain.

Siddharthan, Advaith, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic

simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 896–902, Geneva, Switzerland.

Sridhar, Vivek Kumar Rangarajan, Ani Nenkova, Shrikanth Narayanan, and Dan Jurafsky. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of the 4th Conference on Speech Prosody*, pages 453–456, Campinas, Brazil.

Steinberger, Josef, Massimo Poesio, Mijail Alexandrov Kabadjov, and Karel Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)*, pages 1251–1257, Montreal, Quebec, Canada.

van Deemter, Kees. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

van Deemter, Kees, Albert Gatt, Roger van Gompel, and Emiel Krahmer. 2010. Production of Referring Expressions (PRE-CogSci) 2009: Bridging the gap between computational and empirical approaches to reference. *Journal of Memory and Language*, 54:554–573.

van Halteren, Hans and Simone Teufel. 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.

Viethen, Jette and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 63–70, Sydney, Australia.

Wacholder, Nina, Yael Ravin, and Misook Choi. 1997. Disambigaution of names in text. In *Proceedings of the Fifth Conference on Applied NLP*, pages 202–208, Washington, DC.

Zajic, David, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management, Special Issue on Summarization*, 43:1549–1570.