Proceedings of HLT-NAACL 2003
Short Papers , pp. 70-72
Edmonton, May-June 2003

# References to Named Entities: a Corpus Study

**Ani Nenkova** and **Kathleen McKeown**
Columbia University
Computer Science Department
New York, NY 10027
{ani,kathy}@cs.columbia.edu

## Abstract

References included in multi-document summaries are often problematic. In this paper, we present a corpus study performed to derive a statistical model for the syntactic realization of referential expressions. The interpretation of the probabilistic data helps us gain insight on how extractive summaries can be rewritten in an efficient manner to produce more fluent and easy-to-read text.

## 1 Introduction

Automatically generated summaries, and particularly multi-document summaries, suffer from lack of coherence One explanation is that the most widespread summarization strategy is still sentence extraction, where sentences are extracted word for word from the original documents and are strung together to form a summary. Syntactic form and its influence on summary coherence have not been taken into account in the implementation of a full-fledged summarizer, except in the preliminary work of (Schiffman et al., 2002).

Here we conduct a corpus study focusing on identifying the syntactic properties of first and subsequent mentions of people in newswire text (e.g., "Chief Petty Officer Luis Diaz of the U.S. Coast Guard in Miami" followed by "Diaz"). The resulting statistical model of the flow of referential expressions suggest a set of rewrite rules that can transform the summary back to a more coherent and readable text.

In the following sections, we first describe the corpus that we used and then the statistical model that we developed. It is based on Markov chains and captures how subsequent mentions are conditioned by earlier mentions. We close with discussion of our evaluation, which measures how well the highest probability path in the model can be used to regenerate the sequence of references.

## 2 The Corpus

We used a corpus of news stories, containing 651,000 words drawn from six different newswire agencies, in order to study the syntactic form of noun phrases in which references to people have been realized. We were interested in the occurrence of features such as type and number of premodifiers, presence and type of postmodifiers, and form of name reference for people.

We constructed a large, automatically annotated corpus by merging the output of Charniak's statistical parser (Charniak, 2000) with that of the IBM named entity recognition system Nominator (Wacholder et al., 1997). The corpus contains 6240 references. In this section, we describe the features that were annotated.

Given our focus on references to mentions of people, there are two distinct types of premodifiers, "titles" and "name-external modifiers". The titles are capitalized noun premodifiers that conventionally are recognized as part of the name, such as "president" in "President George W. Bush". Name-external premodifiers are modifiers that do not constitute part of the name, such as "Irish flutist" in "Irish flutist James Galway".

The three major categories of postmodification that we distinguish are apposition, prepositional phrase modification and relative clause. All other postmodifications, such as remarks in parenthesis and verb-initial modifications are lumped in a category "others".

There are three categories of names corresponding to the general European and American name structure. They include full name (first+(middle initial)+last), last name only, and nickname (first or nickname).

In sum, the target NP features that we examined were:

- Is the target named entity the head of the phrase or not? Is it in a possessive construction or not?
- If it is the head, what kind of pre- and post- modification does it have?
- How was the name itself realized in the NP?

In order to identify the appropriate sequences of syntactic forms in coreferring noun phrases, we analyze the coreference chains for each entity mentioned in the text. A coreference chain consists of all the mentions of an entity within a document. In a manually built corpus, a coreference chain can include pronouns and common nouns that refer to the person. However, these forms could not be automatically identified, so coreference chains in our corpus only include noun phrases that contain at least one word from the name. There were 3548 coreference chains in the corpus.

## 3 A Markov Chain Model

The initial examination of the data showed that syntactic forms in coreference chains can be effectively modeled by Markov chains.

Let $X_n$ be random variables taking values in I. We say that $(X_n)_{n\geq 0}$ is a Markov chain with initial distribution $\lambda$ and transition matrix $P$ if

- $X_0$ has distribution $\lambda$
- for $n \geq 0$, conditional on $X_n = i$, $X_{n+1}$ has distribution $(p_{ij}|j \in I)$ and is independent of $X_0, \ldots, X_{n-1}$.

These properties have very visible counterparts in the behavior of coreference chains. The first mention of an entity does have a very special status and its appropriate choice makes text more readable. Thus, the initial distribution of a Markov chain would correspond to the probability of choosing a specific syntactic realization for the first mention of a person in the text. For each subsequent mention, the model assumes that only the form of the immediately preceding mention determines its form. Moreover, the Markov chain model is more informative than other possible approaches to modelling the same phenomena (Nenkova and McKeown, 2003).

| | modification | no modification |
|---|---|---|
| initial | **0.76** | 0.24 |
| modification | 0.44 | 0.56 |
| no modification | 0.24 | **0.75** |

Figure 1: Markov chain for modification transitions. The first row gives the initial distribution vector. $(i, j)$ gives the probability of going from form $i$ to form $j$.

| | full name | last name | nickname |
|---|---|---|---|
| initial | **0.97** | 0.02 | 0.01 |
| full name | 0.20 | **0.75** | 0.05 |
| last name | 0.06 | **0.91** | 0.02 |
| nickname | 0.24 | 0.22 | **0.53** |

Figure 2: Markov chain for name realization. The first row gives the initial distribution vector.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| initial | 0.49 | 0.22 | 0.16 | 0.08 | 0.03 | 0.01 | 0.01 |
| 0 | 0.86 | 0.09 | 0.04 | - | - | - | - |
| 1 | 0.43 | 0.50 | 0.05 | - | - | - | - |
| 2 | 0.78 | 0.13 | 0.08 | - | - | - | - |
| 3 | 0.78 | 0.13 | 0.07 | - | - | - | - |
| 4 | 0.74 | 0.09 | 0.15 | 0.02 | - | - | - |
| 5 | 0.90 | 0.10 | - | - | - | - | - |
| 6 | 0.81 | 0.06 | 0.13 | - | - | - | - |

Figure 3: Markov chain for the number of premodifiers. Probabilities given for merged title and external ones and values below 0.01 are given as dashes.

## 4 Model Interpretation

The number of possible syntactic forms, which corresponds to the possible combination of features, is large, around 160. Because of this, it is not easy to interpret the results if they are taken in their full form. We now show information for one feature at a time so that the tendencies can become clearer.

A first mention is very likely to be modified in *some* way (probability of 0.76, Figure 1), but it is highly unlikely that it will be *both* pre- and postmodified (probability of 0.17). The Markov model predicts that at each next mention, modification can be either used or not, *but* once a non-modified form is chosen, the subsequent realizations will most likely not use modification any more.

From the Markov chain that models the form of names (Figure 2) we can see that first name or nickname mentions are very unlikely. But it also predicts that if such a reference is once chosen, it will most likely continue to be used as a form of reference. This is intuitively very appealing as it models cases where journalists call celebrities by their first name (e.g., "Britney" or "Lady Diana" are often repeatedly used within the same article).

Prepositional, relative clause and "other" modifications appear with equal extremely low probability (in the range 0.01–0.04) after any possible previous mention realization. Thus the syntactic structure of the previous mention cannot be used as a predictor of the appearance of any of these kinds of modifications, so for the task of rewriting references they should not be considered in any way but as "blockers" of further modification. The only type of postmodification with significantly high probability of 0.25 is apposition at the first mention.

Figure 3 shows the probabilities for transitions between NPs with a different number of premodifiers. The mass above the diagonal is almost zero, showing that each subsequent mention has fewer premodifiers than the previous. There are exceptions which are not surprising; for example, a mention with one modifier is usually followed by a mention with one modifier (probability 0.5) accounting for title modifiers such as "Mr." and "Mrs.".

## 5 Rewrite Rules

The Markov chain model derived in the manner described above helps us understand what a typical text looks like. The Markov chain transitions give us defeasible preferences that are true for the average text. Human writers seek more style, so even statistically highly unlikely realizations can be used by a human writer. For example, even a first mention with a pronoun can be felicitous at times. The fact that we were seeking preferences rather than rules allows us to take advantage of the sometimes inaccurate automatically derived corpus. There have inevitably been parser errors or mistakes in Nominator's output, but these can be ignored since, given the large amount of data, the general preferences in realization could be captured even from imperfect data.

We developed a set of rewrite rules that realize the highest probability paths in the Markov chains for name form and modification. In the cases where the name serves as a head of the NP it appears in, the highest probability paths suggest the following:

- **name realization:** use full name at the first mention and last name only at subsequent mentions. The probability of such sequence of transitions is 0.66, compared with 0.01 for last name—full name—last name for example.
- **modification:** the first mention is modified and subsequent mentions are not. As for the type of modification—premodifiers are preferred and in case they cannot be realized, apposition is used. Appositions and premodifiers are removed from any subsequent mention.

The required type of NP realization is currently achived by extracting NPs from the original input documents.

## 6 Evaluation

The rules were used to rewrite 11 summaries produced by the Columbia University summarizer. Four human judges were then given the pairs of the original summary and its rewritten variant (Figure 4). They were asked to decide if they prefer one text over the other or if they are equal. The majority preference was always for the rewritten version and it could be reached in all but one case, where two of the judges preferred the rewritten version and two, the original. The distribution of the 44 individual preferences for a rewritten or original summary were 89% for the rewrite version, 9% for the original version and 2% no preference for a version.

The rewrite module is currently implemented and it runs daily as part of the Columbia Newsblaster summarization system that can be found online at http://newsblaster.cs.columbia.edu.

Figure 4: An example of rewriting references

**Original summary:**
Presidential advisers do not blame **O'Neill**, but they've long recognized that a shakeup of the economic team would help indicate **Bush** was doing everything he could to improve matters. **U.S. President George W. Bush** pushed out **Treasury Secretary Paul O'Neill** and top economic adviser Lawrence Lindsey on Friday, launching the first shake - up of his administration to tackle the ailing economy before the 2004 election campaign.

**Rewritten summary:**
Presidential advisers do not blame **Threasury Secretary Paul O'Neill**, but they've long recognized that a shakeup of the economic team would help indicate **U.S. President George W. Bush** was doing everything he could to improve matters. **Bush** pushed out **O'Neill** and White House economic adviser Lawrence Lindsey on Friday, launching the first shake-up of his administration to tackle the ailing economy before the 2004 election campaign.

## 7 Conclusion and Future work

As has been seen, a major improvement of summary readability can be achieved by using the simple set of rewrite rules that realize the highest probability path in the derived Markov model. One possible usage of the model which is not discussed in the paper but is the focus of current and ongoing work, is to generate realizations "on demand". Referring expressions can be generated by recombining different pieces of the input rather than the currently used extraction of full NPs. This approach will make better use of the Markov model, but it also requires work towards deeper semantic processing of the input. Semantic information is needed in order to prevent the combination of almost synonymous premodifiers in the same NP and also for the identification of properties that are more central for the enity with respect to the focus of the input cluster.

## References


E. Charniak. 2000. A maximum-entropy-inspired parser. In *NAACL-2000.*

A. Nenkova and K. McKeown. 2003 A Corpus Study for Modeling the Syntactic Realization of Entities *Columbia University Tech Report CUCS-001-03*

B. Schiffman, A. Nenkova, and K. McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the HLT'02 Conference.*

N. Wacholder, Y. Ravin, and M. Choi. 1997. Disambigaution of names in text. In *Proceedings of the Fifth Conference on Applied NLP*, pages 202–208.