Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities

Ani Nenkova
Columbia University
ani@cs.columbia.edu

Kathleen McKeown
Columbia University
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 statistical models for the syntactic realization of referential expressions. Our work shows how the syntactic realization of entities can influence the coherence of the text and provides a model for rewriting references in multi-document summaries to smooth disfluencies.

1 Introduction

Automatically generated summaries, and particularly multi-document summaries, suffer from lack of coherence (Boguraev and Neff, 2000). One explanation for this fact 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. While some researchers have developed methods to regenerate summary text from the text of the original articles (e.g., (Barzilay et al., 1999; Jing, 2000; Knight and Marcu, 2002; Schiffman et al., 2001)), the focus has been mostly on removing irrelevant and redundant phrases.

Outside of summarization, though, different aspects of coherence have been studied in great detail. In particular, seminal work on centering (Grosz et al., 1995) motivated numerous investigations of the factors that influence the *local coherence* of discourse. Centering theory looks at two main sources of (in)coherence – the syntactic realization of discourse entities and the transition between focused entities. (Barzilay et al., 2002) have shown how considerations of the latter kind can be used to guide ordering in multi-document summaries. But 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).

Considerations of local coherence are extremely important for summaries, which are very short by definition and thus, are less affected by deficiencies in global discourse structure. Figure 1 shows a summary generated by a fully implemented and extensively used summarizer. The summary gives a good idea of what incoherence problems can arise—the first mention of the two politicians in the summary uses the last names only and this makes it difficult for the reader to know who the summary is referring to. In contrast, fully modified references that identify the entities for the reader occur in the second summary sentence, an odd choice after the bare proper nouns that occurred earlier. The third summary sentence exasperates the problem by using even more modification.

These difficulties of text comprehension due to inappropriate syntactic form have been discussed previously. One of the main claims of centering theory is that different syntactic realizations pose different processing requirements on the hearer and thus contribute to the coherence of discourse. (Krahmer and Theune, 2002) report an experiment on human preferences on sequences of syntactic forms that demonstrates that people prefer subsequent mentions that

Figure 1: A problematic summary

Terrell had 56 percent of the white vote to 31 percent for **Landrieu**, while Landrieu had 75 percent of the black vote to 10 percent for Terrell. A poll released this week shows the race between **Democratic Sen. Mary Landrieu** and **her Republican challenger, Suzanne Haik Terrell**, to be dead even. Voters go to the polls Saturday. With Louisiana's Senate run-off election just four days away, President Bush led the GOP charge Tuesday for **Republican candidate Suzanne Haik Terrell** in what polls now suggest is a toss-up race against **freshman Democratic Sen. Mary Landrieu**.

are less informative than the previous mentions of the same entity. They also cite experiments that show that utterances are more difficult to read if a definite description or a proper name is used in places where pronouns can be appropriate (Gordon et al., 1993).

In this paper, we conduct a corpus study to identify the syntactic properties of first and subsequent mentions of people in newswire text. The resulting statistical model of the flow of referential expressions in text is based on features that can be derived from full text using shallow parsing technology. Thus, it can be used to create a set of recommended rewrite rules that can transform the summary back to a more coherent and readable text. Our study focuses on noun phrases containing mention of people names. These constitute a subset of the general problem of reference in summaries that exemplify under and over-specification in reference. Yet, restriction to people names allows us to build a working solution due to recent advances in language technology, namely statistical parsing and named entity recognition.

In the following sections, we first describe the corpora that we used and then two statistical models that we developed for the task. The first is based on Markov chains and models how subsequent mentions are conditioned by earlier mentions, while the second, stratified model captures the different types of realizations for first through fifth mention separately. We close with discussion of our evaluation, which measures how well the models can regenerate the sequence of references in a test corpus, demonstrating that the Markov model is far more informative.

2 The Corpus

We used a corpus of news, 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 used a variety of sources in order to avoid the possibility of learning paper-specific editor rules from a single source. 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 began our study by manually annotating a small corpus of six articles; this pilot study allowed us to determine which features of interest could be automatically extracted. We then 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 automatically derived 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". (Charniak, 2001), for instance, discusses statistical techniques for disambiguating name structure in examples like the one above, and shows how the structure can be parsed to identify the first name, the last name, middle initial and title modifiers.

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".

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

Examples of the different properties coded for noun phrases are given in Figure 2. In sum, the features of the target NP 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 noun phrase?

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 the 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 the automatically derived corpus only include noun phrases that contain at least one word from the name. There were 3548 coreference chains in the automatically derived corpus; an example is given in Figure 3 which shows both the full coreference derived manually and the abbreviated chain identified in the automatically derived corpus.

Figure 2: Examples of different syntactic forms

NP1: John Aquilino, a former NRA official who now publishes his own gun owners' newsletter.
Codes as: full name + apposition

NP2: Chief Petty Officer Luis Diaz of the U.S. Coast Guard in Miami.
Coded as: 3 title premodifiers + full name + prepositional phrase postmodification

NP3: Dutch speed skater Yvonne van Gennip
Coded as: 3 name-external premods + middle initial name

NP4: Soviet pianist Vladimir Feltsman, who arrived in the United States last August after an eight-year battle to emigrate,
Coded as: 2 name-external + full name + relative clause

NP5: Powell, stationed behind the group of reporters who were questioning Reagan during an Oval Office photo opportunity,
Coded as: last name + other postmodification

3 Statistical Models of Mention Sequence

We developed two models of syntactic realization. The first uses a Markov chain model; it represents the influence of each mention on the subsequent reference. The second models the likelihood of particular forms of syntactic realization for first and subsequent mentions separately. Our results show that the the Markov chain model is more informative than the stratified model.

3.1 The Markov Chain Model

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

The formal definition of a Markov chain follows:

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 λ and transition matrix P if

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

Informally, a Markov chain is given by a transition matrix and an initial distribution. The transition matrix gives the probability of moving from one state to the next, while the initial distribution gives the probability of being in a specific state at time zero. The probability of being in a given state at a given time depends only on the probability of the preceding state at the previous time. All these properties have very visible counterparts in the behavior of coreference chains. The first mention of an entity does have a very special status ((Fraurud, 1990) and (Poesio and Vieira, 1998)) 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. This property of the model will be tested in evaluation, but seems to predict intuitively plausible sequences. For example, if a person has been previously mentioned by full name, then it

Figure 3: Full coreference chain for a person. The number in parenthesis shows how many consecutive times the given syntactic form has been repeated in the chain. The entities marked with an "X" represent the chain that will be derived automatically

Bill Clinton X
the newly installed President
The man, whom almost two-thirds of all Americans trusted
Bill Clinton X
Clinton X
he (2)
Clinton X
he
Clinton (5) X
Bill Clinton X
Clinton (2) X
the president
Clinton (2) X
Bill Clinton X
the President

is most likely appropriate to refer to him again by his last name, while if the previous mention was a first name, then a subsequent first name mention is appropriate.

Of course, additional discourse factors can play a role in determining the use of a given type of NP. For example, (Fox, 1998) and (Levy, 1984) give detailed studies of the global context factors that play a role in syntactic realization. But for now, we will adopt the simple Markov chain model to see how useful it can be. The hope is that, at a later stage, a more sophisticated account of context can augment the model and make it more reliable.

3.2 The Stratified Model

The stratified model is guided by the idea that it is not just the first mention that has special characteristics, but rather that there are special features of the syntactic form strongly associated with the first mention, other features associated with the second mention and so on. Since summaries are short, we can safely make the assumption that a person will not be mentioned more than five times; thus, the model will look for features of the syntactic realizations from the first up to the fifth mention of an entity.

For each $m = 1, ..., 5$ we compute the probability

$$P(SF|m) = \frac{P(SF,m)}{P(m)} = \frac{\frac{cnt(SF,m)}{total}}{\frac{cnt(m)}{total}} = \frac{cnt(SF,m)}{cnt(m)},$$

where SF is the variable corresponding to the syntactic realization, m is the number of mention and $total$ is the number of syntactic forms counted for the entire corpus.

3.2.1 Model Comparison

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.

Figure 4 shows that a first mention is very likely to be modified in *some* way (probability of 0.76), but it is highly unlikely that it will be *both* premodified and postmodified (probability of 0.17).

The results for the presence or absence of modification of some kind are given in Figures 4 and 8. The stratified model does not tell us anything about whether a subsequent mention should be modified; both cases are almost equally likely for mentions from the second up to the fifth. The Markov chain gives us more useful information – it 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.

The Markov chain that models the form of names (Figure 5) also gives us more information than the stratified model (Figure 8). From the latter we can see that first name or nickname mentions are very unlikely. But the Markov chain 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 commonly used while "Spears" or "Spencer" are not).

Also, the analysis of the data leads us to reject the hypothesis that there are some specific features of *each* number of mention. The distribution of features for mentions after the first are almost identical. This is one more reason to give preference to the Markov model over the stratified one, since the Markov model gives special importance to the first mention and treats all subsequent mentions equally.

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

Figure 4: Markov chain for modification transitions. The first row gives the initial distribution vector.

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

Figure 6 shows the probabilities of transitions between the different kinds of postmodification. Prepositional, relative clause and "other" modifications appear with equal extremely low probability 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 (see Section 4) they will not be considered in any way but as "blockers" of further modification.

Figure 7 shows the probabilities for transitions between NPs with a different number of premodifiers. It can be seen that the mass above the diagonal is close to zero, which means that each subsequent mention has fewer premodifiers than the previous one. There are some exceptions to this rule which are not surprising; for example, a mention with one modifier is usually followed by a mention with one modifier (probability 0.5) since title modifiers such as "Mr.", "Mrs." are included in the counts for transitions. There are newspaper specific rules about the usage of these modifiers. The Wall Street Journal and New York Times, for example, do use them as a rule, except for historical and criminal figures. Such editor rules are interesting and can be useful for summarization, but the information that can be gathered from them is too subtle to encode computationally. As can be seen in the later section, these honorifics will be treated in rewrite as any other premodifier and will be dropped at subsequent mention.

We also looked separately at the cases when the first mention of a person was not the head of the noun phrase (e.g., "the Bush administration", "Mendeleev's periodic table"). Such name mentions obviously do not have postmodification of any kind and premodification cannot be reliably identified automatically since current parsers output flat structure for noun premodifiers. Words in the NP that precede the name can either modify the name or the head, so no conclusion can be drawn. The only relevant feature in this case was the name realization. The model built for those entities whose first mentions are in a non-head NP differs quite a bit from the model for head NPs. For example, full names are used in first mentions in non-head, non-possessive constructions in only 75% of all first mentions. This was the reason why in the rewrite rules that we developed and discuss below, the name is not changed in any way if its first mention is in a non-head position.

4 What was learned

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 as can be seen in Figure 9. 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,

	apposition	none	prepositional	relcl	other
initial	**0.25**	0.60	*0.07*	*0.04*	*0.04*
apposition	0.06	0.88	*0.00*	*0.04*	*0.02*
none	0.04	0.89	*0.01*	*0.03*	*0.03*
prepositional	0.10	0.80	*0.01*	*0.07*	*0.02*
relcl	0.08	0.82	*0.01*	*0.06*	*0.03*
other	0.07	0.88	*0.00*	*0.04*	*0.01*

Figure 6: Markov chain for postmodification.

	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*	*0.01*	*0.00*	*0.00*	*0.00*
1	0.43	0.50	*0.05*	*0.01*	*0.00*	*0.01*	*0.00*
2	0.78	0.13	0.08	*0.01*	*0.01*	*0.00*	*0.00*
3	0.78	0.13	0.07	0.01	*0.01*	*0.00*	*0.00*
4	0.74	0.09	0.15	0.02	0.00	*0.00*	*0.00*
5	0.90	0.10	0.00	0.00	0.00	0.00	*0.00*
6	0.81	0.06	0.13	0.00	0.00	0.00	0.00

Figure 7: Markov chain for the number of premodifiers. Count given for merged title and external premodifiers.

the general preferences in realization could be captured even from imperfect data.

Figure 9: A first paragraph from our hand-annotated corpus. A first mention by pronoun is possible, but highly unlikely

He moved into the governor's mansion at 32, **heir to a tradition of progressive Southern governors** and ready to light up Arkansas. It was January 1979 and there was so much to do: Education needed to be overhauled, the business climate needed to be improved, the state needed to be dragged out of its slumberous, defeatist past. **Bill Clinton, the youngest governor in the nation since Harold Stassen,** had such big plans.

Since summaries are generated by a computer and not a human, deviation from the standard preference can very likely introduce a problem in the summary rather than make it more stylish. Thus the learned defeasible preferences help us decide when a reference in a summary needs to be rewritten and also it suggests the type of rewrite needed.

We developed a set of rewrite rules through manual analysis of the Markov chain model. This is a subset of the full power of the model, but it dramatically improves the quality of references.

1. For first mentions
   (a) If the person's name is the head of the noun phrase,
      i. If a person is mentioned by last name, insert full name and the longest, in number of words, premodifier found in the input articles. The first mention from the article from which the summary sentence is drawn is preferred.
      ii. If no premodification is found in the input, check all first mentions in the input to see if any of them includes an apposition modifier. Take the longest such modifier and include it in the first mention NP.
   (b) The name is not modified at all if it is not the head of the noun phrase it appears in.
2. For all subsequent mentions use last name only, remove all premodifiers and delete all apposition modifiers.

	first	second	third	foutrh	fifth
modified	0.76	0.48	0.52	0.54	0.51
non-modified	0.24	0.52	0.48	0.46	0.49
premodified	0.51	0.37	0.42	0.45	0.43
non-premodified	0.49	0.63	0.58	0.55	0.57
full name	0.97	0.13	0.12	0.10	0.10
last name	0.02	0.81	0.82	0.84	0.83
nickname	0.01	0.05	0.06	0.06	0.07

Figure 8: Probabilities of an NP being modified and non-modified at a particular mention.

The above straightforward rules lead to the rewritten version of the summary in Figure 10.

Figure 10: Rewritten summary

Republican candidate Suzanne Haik Terrell had 56 percent of the white vote to 31 percent for **Democratic Sen. Mary Landrieu**, while **Landrieu** had 75 percent of the black vote to 10 percent for **Terrell**. A poll released this week shows the race between **Landrieu** and **her Republican challenger, Terrell,** to be dead even. Voters go to the polls Saturday. Emboldened by November election triumphs, President Bush urged Louisiana voters on Tuesday to pad the GOP Senate majority and defeat a Democratic incumbent who claims her own Bush-friendly voting record. With Louisiana's Senate run-off election just four days away, Bush led the GOP charge Tuesday for **Terrell** in what polls now suggest is a toss - up race against **Landrieu**.

5 Evaluation

The above three rules were used to rewrite 11 summaries produced by our summarizer. Four human judges were then given the pairs of the original summary and its rewritten variant. 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 preferences is shown on Figure 11.

In only one case a majority preference could not be reached, with two of the judges preferring the rewritten version and two, the original. This particular summary was controversial because it included non-name references to people, such as "the president" in the first coreference chain shown on Figure 3 . This type of coreference could not be identified in our automatic approach and thus, the fact of its occurrence was not taken into account during rewrite. This shows that work on person centered coreference can be very helpful for summarization as well.

There were two more cases where one judge showed preference for the original version. They both came with comments that the reason for the preference was that the original version exhibited more variation. Thus, it seems that the rule for strictly using last name at subsequent mentions is too rigid and most probably will need modification in cases where a person is mentioned more than three times.

6 Conclusions and Future Work

We have shown how simple syntactic considerations can improve a multi-document summary by making it more coherent. A Markov model for transitions between syntactic realizations was derived and used for composing initial rewrite rules. This approach to summarization, focusing on summary revision, has not been used in the area so far. Existing summarization approaches that do make changes in the sentences extracted from the original input have as a goal the reduction of information/number of words, while in our approach the coherence and readability of the summary are of primary consideration.

As can be seen, a major improvement can be achieved even by using the proposed simple set of rewrite rules used for the evaluation. But they do not fully reflect all we learned from the data. These rules will be expanded with the rule for nickname usage discussed above. The rule for dropping premodification on subsequent mentions will also be refined so that it takes into account the gradual shrinking in the number of premodifiers. In order to do this, we will need to build some kind of simple discourse model

rewrite version	original version	none
89%	9%	2%

Figure 11: Distribution of the 44 individual preferences for a rewritten or original summary

so that within it we can track which properties of an entity have already been realized and which can be realized in subsequent mentions.

One possible usage of the Markov model not discussed in the paper is to generate realizations "on demand" so that the highest probability path in the model can be realized in the summary. This means that referring expressions will be generated by recombining different pieces of the input rather then the currently used extraction of full NPs. For this task, again, a discourse model will be needed and the information in it will be used as a knowledge base for the generation process.

Another direction of immediate future work will be to explore the possibilities of applying the same approach to common nouns. Reference realization for common nouns is more complex than for people and it will be interesting to see how the work in NLG in referring expression generation can be adapted for the task of multiple document summarization.

References

R. Barzilay, K. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

R. Barzilay, N. Elhadad, and K. McKeown. 2002. Inferring strategies for sentence ordering in multi-document summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

B. Boguraev and M. Neff. 2000. Lexical cohesion, discourse segmentation and document summarization. In *RIAO-2000, Content-Based Multimedia Information Access*.

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

E. Charniak. 2001. Unsupervised learning of name structure from coreference data. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 48–54.

B. Fox. 1998. *Discourse structure and anaphora*. Cambridge University Press.

K. Fraurud. 1990. Definiteness and the processing of nps in discourse. *Journal of Semantics*, 7:395–433.

P. Gordon, B. Grosz, and L. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, (17):311–347.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

H Jing. 2000. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied NLP Conference, ANLP'2000*.

K. Knight and D. Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1).

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications.

E. Levy. 1984. *Communicating Thematic Structure: The Use of Referring Terms and Gestures in Narrative Discourse*. PhD thesis.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

B. Schiffman, Inderjeet. Mani, and K. Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, Toulouse, France, July.

B. Schiffman, A. Nenkova, and K. McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

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