Entity-driven Rewrite for Multi-document Summarization

Ani Nenkova
University of Pennsylvania
Department of Computer and Information Science
nenkova@seas.upenn.edu

Abstract

In this paper we explore the benefits from and shortcomings of entity-driven noun phrase rewriting for multi-document summarization of news. The approach leads to 20% to 50% different content in the summary in comparison to an extractive summary produced using the same underlying approach, showing the promise the technique has to offer. In addition, summaries produced using entity-driven rewrite have higher linguistic quality than a comparison non-extractive system. Some improvement is also seen in content selection over extractive summarization as measured by pyramid method evaluation.

1 Introduction

Two of the key components of effective summarizations are the ability to identify important points in the text and to adequately reword the original text in order to convey these points. Automatic text summarization approaches have offered reasonably well-performing approximations for identifiying important sentences (Lin and Hovy, 2002; Schiffman et al., 2002; Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Daumé III and Marcu, 2006) but, not surprisingly, text (re)generation has been a major challange despite some work on sub-sentential modification (Jing and McKeown, 2000; Knight and Marcu, 2000; Barzilay and McKeown, 2005). An additional drawback of extractive approaches is that estimates for the importance of larger text units such as sentences depend on the length of the sentence (Nenkova et al., 2006).

Sentence simplification or compaction algorithms are driven mainly by grammaticality considerations. Whether approaches for estimating importance can be applied to units smaller than sentences and used in text rewrite in the summary production is a question that remains unanswered. The option to operate on smaller units, which can be mixed and matched from the input to give novel combinations in the summary, offers several possible advantages.

Improve content Sometimes sentences in the input can contain both information that is very appropriate to include in a summary and information that should not appear in a summary. Being able to remove unnecessary parts can free up space for better content. Similarly, a sentence might be good overall, but could be further improved if more details about an entity or event are added in. Overall, a summarizer capable of operating on subsentential units would in principle be better at content selection.

Improve readability Linguistic quality evaluation of automatic summaries in the Document Understanding Conference reveals that summarizers perform rather poorly on several readability aspects, including referential clarity. The gap between human and automatic performance is much larger for linguistic quality aspects than for content selection. In more than half of the automatic summaries there were entities for which it was not clear what/who they were and how they were related to the story. The ability to add in descriptions for entities in the summaries could improve the referential clarity of summaries and can be achieved through text rewrite

of subsentential units.

IP issues Another very practical reason to be interested in altering the original wording of sentences in summaries in a news browsing system involves intellectual property issues. Newspapers are not willing to allow verbatim usage of long passages of their articles on commercial websites. Being able to change the original wording can thus allow companies to include longer than one sentence summaries, which would increase user satisfaction (McKeown et al., 2005).

These considerations serve as direct motivation for exploring how a simple but effective summarizer framework can accommodate noun phrase rewrite in multi-document summarization of news. The idea is for each sentence in a summary to automatically examine the noun phrases in it and decide if a different noun phrase is more informative and should be included in the sentence in place of the original. Consider the following example:

Sentence 1 *The arrest* caused an international controversy.

Sentence 2 *The arrest in London of former Chilean dictator Augusto Pinochet* caused an international controversy.

Now, consider the situation where we need to express in a summary that the arrest was controversial and this is the first sentence in the summary, and sentence 1 is available in the input ("The arrest caused an international controversy"), as well as an unrelated sentence such as "The arrest in London of former Chilean dictator Augusto Pinochet was widely discussed in the British press". NP rewrite can allow us to form the rewritten sentence 2, which would be a much more informative first sentence for the summary: "The arrest in London of former Chilean dictator Augusto Pinochet caused an international controversy". Similarly, if sentence 2 is available in the input and it is selected in the summary after a sentence that expresses the fact that the arrest took place, it will be more appropriate to rewrite sentence 2 into sentence 1 for inclusion in the summary.

This example shows the potential power of noun phrase rewrite. It also suggests that context will play a role in the rewrite process, since different noun phrase realizations will be most appropriate depending on what has been said in the summary up to the point at which rewrite takes place.

2 NP-rewrite enhanced frequency summarizer

Frequency and frequency-related measures of importance have been traditionally used in text summarization as indicators of importance (Luhn, 1958; Lin and Hovy, 2000; Conroy et al., 2006). Notably, a greedy frequency-driven approach leads to very good results in content selection (Nenkova et al., 2006). In this approach sentence importance is measured as a function of the frequency in the input of the content words in that sentence. The most important sentence is selected, the weight of words in it are adjusted, and sentence weights are recomputed for the new weights beofre selecting the next sentence.

This conceptually simple summarization approach can readily be extended to include NP rewrite and allow us to examine the effect of rewrite capabilities on overall content selection and readability. The specific algorithm for frequency-driven summarization and rewrite is as follows:

Step 1 Estimate the importance of each content word w_i based on its frequency in the input n_i, $p(w_i) = \frac{n_i}{N}$.

Step 2 For each sentence S_j in the input, estimate its importance based on the words in the sentence $w_i \in S_j$: the weight of the sentence is equal to the average weight of content words appearing in it.

$$Weight(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|w_i \in S_j|}$$

Step 3 Select the sentence with the highest weight.

Step 4 For each maximum noun phrase NP_k in the selected sentence

4.1 For each coreferring noun phrase NP_i, such that $NP_i \equiv NP_k$ from all input documents, compute a weight $Weight(NP_i) = F_{RW}(w_r \in NP_i)$.

4.2 Select the noun phrase with the highest weight and insert it in the sentence in

place of the original NP. In case of ties, select the shorter noun phrase.

Step 5 For each content word in the rewritten sentence, update its weight by setting it to 0.

Step 6 If the desired summary length has not been reached, go to step 2.

Step 4 is the NP rewriting step. The function F_{RW} is the rewrite composition function that assigns weights to noun phrases based on the importance of words that appear in the noun phrase. The two options that we explore here are $F_{RW} \equiv Avr$ and $F_{RW} \equiv Sum$; the weight of an NP equals the average weight or sum of weights of content words in the NP respectively. The two selections lead to different behavior in rewrite. $F_{RW} \equiv Avr$ will generally prefer the shorter noun phrases, typically consisting of just the noun phrase head and it will overall tend to reduce the selected sentence. $F_{RW} \equiv Sum$ will behave quite differently: it will insert relevant information that has not been conveyed by the summary so far (add a longer noun phrase) and will reduce the NP if the words in it already appear in the summary. This means that $F_{RW} \equiv Sum$ will have the behavior close to what we expect for entity-centric rewrite: inluding more descriptive information at the first mention of the entity, and using shorter references at subsequent mentions.

Maximum noun phrases are the unit on which NP rewrite operates. They are defined in a dependency parse tree as the subtree that has as a root a noun such that there is no other noun on the path between it and the root of the tree. For example , there are two maximum NPs, with heads "police" and "Augusto_Pinochet" in the sentence "British police arrested former Chilean dictator Augusto Pinochet". The noun phrase "former chilean dictator" is not a maximum NP, since there is a noun (augusto_pinochet) on the path in the dependency tree between the noun "dictator" and the root of the tree. By definition a maximum NP includes all nominal and adjectival premodifiers of the head, as well as postmodifiers such as prepositional phrases, appositions, and relative clauses. This means that maximum NPs can be rather complex, covering a wide range of production rules in a context-free grammar. The dependency tree definition of maximum noun phrase makes it easy to see why these are a good unit for subsentential rewrite: the subtree that has the head of the NP as a root contains only modifiers of the head, and by rewriting the noun phrase, the amount of information expressed about the head entity can be varied.

In our implementation, a context free grammar probabilistic parser (Charniak, 2000) was used to parse the input. The maximum noun phrases were identified by finding sequences of <*np*>...</*np*> tags in the parse such that the number of opening and closing tags is equal. Each NP identified by such tag spans was considered as a candidate for rewrite.

Coreference classes A coreference class CR_m is the class of all maximum noun phrases in the input that refer to the same entity E_m. The general problem of coreference resolution is hard, and is even more complicated for the multi-document summarization case, in which cross-document resolution needs to be performed. Here we make a simplifying assumption, stating that all noun phrases that have the same noun as a head belong to the same coreference class. While we expected that this assumption would lead to some wrong decisions, we also suspected that in most common summarization scenarios, even if there are more than one entities expressed with the same noun, only one of them would be the *main* focus for the news story and will appear more often across input sentences. References to such main entities will be likely to be picked in a sentence for inclusion in the summary by chance more often than other competeing entities. We thus used the head noun equivalance to form the classes. A post-evaluation inspection of the summaries confirmed that our assumption was correct and there were only a small number of errors in the rewritten summaries that were due to coreference errors, which were greatly outnumbered by parsing errors for example. In a future evaluation, we will evaluate the rewrite module assuming perfect coreference and parsing, in order to see the impact of the core NP-rewrite approach itself.

3 NP rewrite evaluation

The NP rewrite summarization algorithm was applied to the 50 test sets for generic multi-document

summarization from the 2004 Document Understanding Conference. Two examples of its operation with $F_{RW} \equiv Avr$ are shown below.

Original.1 While the British government defended *the arrest*, it took no stand on extradition of Pinochet to Spain.

NP-Rewite.1 While the British government defended *the arrest in London of former Chilean dictator Augusto Pinochet*, it took no stand on extradition of Pinochet to Spain.

Original.2 *Duisenberg* has said growth in the euro area countries next year will be about 2.5 percent, lower than *the 3 percent* predicted earlier.

NP-Rewrite.2 *Wim Duisenberg, the head of the new European Central Bank,* has said growth in the euro area will be about 2.5 percent, lower than *just 1 percent in the euro-zone unemployment* predicted earlier.

We can see that in both cases, the NP rewrite pasted into the sentence important additional information. But in the second example we also see an error that was caused by the simplifying assumption for the creation of the coreference classes according to which the percentage of unemployment and growth have been put in the same class.

In order to estimate how much the summary is changed because of the use of the NP rewrite, we computed the unigram overlap between the original extractive summary and the NP-rewrite summary. As expected, $F_{FW} \equiv Sum$ leads to bigger changes and on average the rewritten summaries contained only 54% of the unigrams from the extractive summaries; for $F_{RW} \equiv Avr$, there was a smaller change between the extractive and the rewritten summary, with 79% of the unigrams being the same between the two summaries.

3.1 Linguistic quality evaluation

Noun phrase rewrite has the potential to improve the referential clarity of summaries, by inserting in the sentences more information about entities when such is available. It is of interest to see how the rewrite version of the summarizer would compare to the extractive version, as well as how its linguistic quality compares to that of other summarizers that participated in DUC. Four summarizers were evaluated: peer 117, which was a system that used generation techniques to produce the summary and was the only real non-extractive summarizer participant at DUC 2004 (Vanderwende et al., 2004); the extractive frequency summarizer, and the two versions of the rewrite algorithm (Sum and Avr). The evaluated rewritten summaries had potential errors coming from different sources, such as coreference resolution, parsing errors, sentence splitting errors, as well as errors coming directly from rewrite, in which an unsuitable NP is chosen to be included in the summary. Improvements in parsing for example could lead to better overall rewrite results, but we evaluated the output as is, in order to see what is the performance that can be expected in a realistic setting for fully automatic rewrite.

SYSTEM	$\mathbf{Q}_1$	$\mathbf{Q}_2$	$\mathbf{Q}_3$	$\mathbf{Q}_4$	$\mathbf{Q}_5$
SUM$_{Id}$	4.06	4.12	3.80	3.80	3.20
SUM$_{Avr}$	3.40	3.90	3.36	3.52	2.80
SUM$_{Sum}$	2.96	3.34	3.30	3.48	2.80
peer 117	2.06	3.08	2.42	3.12	2.10

Table 1: Linguistic quality evaluation. Peer 117 was the only non-extractive system entry in DUC 2004; SUM$_{Id}$ is the frequency summarizer with no NP rewrite; and the two versions of rewrite with sum and average as combination functions.

The evaluation was done by five native English speakers, using the five DUC linguistic quality questions on grammaticality (Q_1), repetition (Q_2), referential clarity (Q_3), focus (Q_4) and coherence (Q_5). Five evaluators were used so that possible idiosyncratic preference of a single evaluator could be avoided. Each evaluator evaluated all five summaries for each test set, presented in a random order. The results are shown in table 3.1. Each summary was evaluated for each of the properties on a scale from 1 to 5, with 5 being very good with respect to the quality and 1, very bad.

Comparing NP rewrite to extraction Here we would be interested in comparing the extractive frequency summarizer (SUM$_{Id}$), and the two version of systems that rewrite noun phrases: SUM$_{Avr}$ (which changes about 20% of the text) and SUM$_{Sum}$ (which changes about 50% of the text). The general trend that we see for all five dimensions of linguistic quality is that the more the text is automatically altered, the worse the linguistic quality of the summary

gets. In particular, the grammaticality of the summaries drops significantly for the rewrite systems. The increase of repetition is also significant between SUM_{Id} and SUM_{Sum}. Error analysis showed that sometimes increased repetition occurred in the process of rewrite for the following reason: the context weight update for words is done only after each noun phrase in the sentence has been rewritten. Occasionally, this led to a situation in which a noun phrase was augmented with information that was expressed later in the original sentence. The referential clarity of rewritten summaries also drops significantly, which is a rather disappointing result, since one of the motivations for doing noun phrase rewrite was the desire to improve referential clarity by adding information where such is necessary. One of the problems here is that it is almost impossible for human evaluators to ignore grammatical errors when judging referential clarity. Grammatical errors decrease the overall readability and a summary that is given a lower grammaticality score tends to also receive lower referential clarity score. This fact of quality perception is a real challenge for summarization systems that move towards abstraction and alter the original wording of sentences since certainly automatic approaches are likely to introduce ingrammaticalities.

Comparing SUM_{Sum} and peer 117 We now turn to the comparison of between SUM_{Sum} and the generation based system 117. This system is unique among the DUC 2004 systems, and the only one that year that experimented with generation techniques for summarization. System 117 is verb-driven: it analizes the input in terms of predicate-argument triples and identifies the most important triples. These are then verbalized by a generation system originally developed as a realization component in a machine translation engine. As a result, peer 117 possibly made even more changes to the original text then the NP-rewrite system. The results of the comparison are consistent with the observation that the more changes are made to the original sentences, the more the readability of summaries decreases. SUM_{Sum} is significantly better than peer 117 on all five readability aspects, with notable difference in the grammaticality and referential quality, for which SUM_{Sum} outperforms peer 117 by a full point. This indicates that NPs are a good candidate granularity for sentence changes and it can lead to substantial altering of the text while preserving significantly better overall readability.

3.2 Content selection evaluation

We now examine the question of how the content in the summaries changed due to the NP-rewrite, since improving content selection was the other motivation for exploring rewrite. In particular, we are interested in the change in content selection between SUM_{Sum} and SUM_{Id} (the extractive version of the summarizer). We use SUM_{Sum} for the comparison because it led to bigger changes in the summary text compared to the purely extractive version. We used the pyramid evaluation method: four human summaries for each input were manually analyzed to identify shared *content units*. The weight of each content unit is equal to the number of model summaries that express it. The pyramid score of an automatic summary is equal to the weight of the content units expressed in the summary divided by the weight of an ideally informative summary of the same length (the content unit identification is again done manually by an annotator).

Of the 50 test sets, there were 22 sets in which the NP-rewritten version had lower pyramid scores than the extractive version of the summary, 23 sets in which the rewritten summaries had better scores, and 5 sets in which the rewritten and extractive summaries had exactly the same scores. So we see that in half of the cases the NP-rewrite actually improved the content of the summary. The summarizer version that uses NP-rewrite has overall better content selection performance than the purely extractive system. The original pyramid score increased from 0.4039 to 0.4169 for the version with rewrite. This improvement is not significant, but shows a trend in the expected direction of improvement.

The lack of significance in the improvement is due to large variation in performance: when np rewrite worked as expected, content selection improved. But on occasions when errors occurred, both readability and content selection were noticeably compromised. Here is an example of summaries for the same input in which the NP-rewritten version had better content. After each summary, we list the content units from the pyramid content analysis that were expressed in the summary. The weight of each

content unit is given in brackets before the label of the unit and content units that differ between the extractive and rewritten version are displayed in italic. The rewritten version conveys high weight content units that do not appear in the extractive version, with weights 4 (maximum weight here) and 3 respectively.

Extractive summary Italy's Communist Refounding Party rejected Prime Minister Prodi's proposed 1999 budget. By one vote, Premier Romano Prodi's center-left coalition lost a confidence vote in the Chamber of Deputies Friday, and he went to the presidential palace to rsign. Three days after the collapse of Premier Romano Prodi's center-left government, Italy's president began calling in political leaders Monday to try to reach a consensus on a new government. Prodi has said he would call a confidence vote if he lost the Communists' support." I have always acted with coherence," Prodi said before a morning meeting with President Oscar Luigi.

(4) Prodi lost a confidence vote

(4) The Refounding Party is Italy's Communist Party

(4) The Refounding Party rejected the government's budget

(3) The dispute is over the 1999 budget

(2) Prodi's coalition was center-left coalition

(2) The confidence vote was lost by only 1 vote

(1) Prodi is the Italian Prime Minister

(1) Prodi wants a confidence vote from Parliament

NP-rewrite version Communist Refounding, a fringe group of hard-line leftists who broke with the minstream Communists after they overhauled the party following the collapse of Communism in Eastern Europe rejected Prime Minister Prodi's proposed 1999 budget. By only one vote, the center-left prime minister of Italy, Romano Prodi, lost The vote in the lower chamber of Parliament 313 against the confidence motion brought by the government to 312 in favor in Parliament Friday and was toppled from power. President Oscar Luigi Scalfaro, who asked him to stay on as caretaker premier while the head of state decides whether to call elections.

(4) Prodi lost a confidence vote

(4) Prodi will stay as caretaker until a new government is formed

(4) The Refounding Party is Italy's Communist Party

(4) The Refounding Party rejected the government's budget

(3) Scalfaro must decide whether to hold new elections

(3) The dispute is over the 1999 budget

(2) Prodi's coalition was center-left coalition

(2) The confidence vote was lost by only 1 vote

(1) Prodi is the Italian Prime Minister

Below is another example, showing the worse deterioration of the rewritten summary compared to the extractive one, both in terms of grammaticality and content. Here, the problem with repetition during rewrite arises: the same person is mentioned twice in the sentence and at both places the same overly long description is selected during rewrie, rendering the sentence practically unreadable.

Extractive summary Police said Henderson and McKinney lured Shepard from the bar by saying they too were gay and one of their girlfriends said Shepard had embarrassed one of the men by making a pass at him. 1,000 people mourned Matthew Shepherd, the gay University of Wyoming student who was severely beaten and left to die tied to a fence. With passersby spontaneously joining the protest group, two women held another sign that read," No Hate Crimes in Wyoming." Two candlelight vigils were held Sunday night. Russell Anderson, 21, and Aaron McKinney, 21, were charged with attempted murder.

(4) The victim was a student at the University of Wyoming

(4) The victim was brutally beaten

(4) The victim was openly gay

(3) The crime was widely denounced

(3) The nearly lifeless body was tied to a fence

(3) The victim died

(3) The victim was left to die

(2) The men were arrested on charges of kidnapping and attempted first degree murder

(2) There were candlelight vigils in support for the victim

(1) Russell Henderson and Aaron McKinney are the names of the people responsible for the death

NP-rewrite version Police said Henderson and McKinney lured the The slight, soft-spoken 21-year-old Shepard, a freshman at the University of Wyoming, who became an overnight symbol of anti-gay violence after he was found dangling from the fence by a passerby from a bar by saying they too were gay and one of their girlfriends said the The slight, soft-spoken 21-year-old Shepard, a freshman at the University of Wyoming, who became an overnight symbol of anti-gay violence after he was found dangling from the fence by a passerby had embarrassed one of the new ads in that supposedly hate-free crusade.

(4) The victim was a student at the University of Wyoming

(3)The nearly lifeless body was tied to a fence *(1) A passerby found the victim*

(1) Russell Henderson and Aaron McKinney are the names of the people responsible for the death

(1) The victim was 22-year old

Even from this unsuccessful attempt for rewrite we can see how changes of the original text can be desirable, since some of the newly introduced information is in fact suitable for the summary.

4 Conclusions

We have demonstrated that an entity-driven approach to rewrite in multi-document summarization can lead to considerably different summary, in terms of content, compared to the extractive version of the same system. Indeed, the difference leads to some improvement measurable in terms of pyramid method evaluation. The approach also significantly outperforms in linguistic quality a non-extractive event-centric system.

Results also show that in terms of linguistic quality, extractive systems will be curently superior to systems that alter the original wording from the input. Sadly, extractive and abstractive systems are evaluated together and compared against each other, putting pressure on system developers and preventing them from fully exploring the strengths of generation techniques. It seems that if researchers in the field are to explore non-extractive methods, they would need to compare their systems separately from extractive systems, at least in the beginning exploration stages. The development of non-extractive approaches in absolutely necessary if automatic summarization were to achieve levels of performance close to human, given the highly abstractive form of summaries written by people.

Results also indicate that both extractive and non-extractive systems perform rather poorly in terms of the focus and coherence of the summaries that they produce, identifying macro content planning as an important area for summarization.

References

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

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

John Conroy, Judith Schlesinger, and Dianne O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of ACL, companion volume*.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Sydney, Australia.

Gunes Erkan and Dragomir Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Hongyan Jing 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)*.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization — step one: Sentence compression. In *Proceeding of The American Association for Artificial Intelligence Conference (AAAI-2000)*, pages 703–710.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501.

Chin-Yew Lin and Eduard Hovy. 2002. Automated multi-document summarization in neats. In *Proceedings of the Human Language Technology Conference (HLT2002)*.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

K. McKeown, R. Passonneau, D. Elson, A. Nenkova, and J. Hirschberg. 2005. Do summaries help? a task-based evaluation of multi-document summarization. In *SIGIR*.

R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP 2004*, pages 404–411.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of SIGIR*.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

Lucy Vanderwende, Michele Banko, and Arul Menezes. 2004. Event-centric summary generation. In *Proceedings of the Document Understanding Conference (DUC'04)*.