Clustering and ordering of sentences for news summarization

CIS 530 Project Progress Report

Quan Dong, Fabian Peternek, Yayang Tian
qdong@seas.upenn.edu, *fape@seas.upenn.edu*, *yaytian@seas.upenn.edu*
December 7, 2010

1 Introduction

The following report will summarize the current progress of our project. It contains two parts: First we detail the work that we have done on the project thus far. Secondly we discuss some related work for our project.

2 Work so far

We are working on sentence ordering for multi-document summarization using approaches found in (Barzilay et al., 2002). Specifically we are implementing the chronological ordering and the augmented algorithm. The original plan was furthermore to modify the augmented algorithm such that it can be used to order sentences instead of clusters of sentences. This however proved to be impossible without considerable changes in the algorithm, as it depends highly on a notion of relatedness that needs clusters to be meaningful.

We therefore decided to use *CLUTO* to cluster the all the sentences in the collection of documents, and then use the mentioned ordering algorithms. After ordering the Clusters in this way, we will use the notion of topic weight to choose one sentence from every cluster to include in the summary.

So far we did the following to achieve this goal: We installed CLUTO and made ourselves familiar with how it works. We furthermore implemented most of the functions, that we will use for the augmented algorithm, however we did not yet put those functions together into an implementation of the full algorithm.

We are still planning to implement a modification of the augmented algorithm in an effort to improve it. However this modification will likely be of fairly small scope. Strictly speaking we are already modifying it, as the sentence selection process (choosing which sentence to take from a cluster) was not automated in the original paper. This is the part we will focus on: Instead of just using topic weight we would like to consider choosing a sentence that has topic words in highly salient positions, for example as a noun.

3 Related work

In this section we will discuss three more papers that have a different approach to the ordering of sentences than the one we are implementing.

3.1 A bottom-up strategy

This section discusses the paper "A Bottom-up Approach to Sentence Ordering for Multi-document Summarization" (Bollegala et al., 2010).

Problem definition The idea here is to combine several ordering strategies that have been developed over the years, to achieve a more coherent summary than one driven from a single method.

Main approaches First an association strength between two segments of text is defined. This is a kind of probability of segment A preceding segment B. Based on related other papers four criteria of determining the association of two segments are defined:

Chronology criterion Reflects the chronological ordering of two segments. The chronological order of arranging segment B after A is determined by the comparison between the last sentence in the segment A and the first sentence in the segment B based on their publishing date.

Topical-closeness criterion Reflects the topic similarity of two segments. The association strength gets a higher value when the topic referred by segment B is similar to that in segment A.

Precedence criterion Reflects the substitutability of the presuppositional information of segment B (e.g. the sentences appearing before sentence in B) as segment A.

Succession criterion Reflects the substitutability of the succeeding information of segment B (e.g. the sentences appearing after sentence in B) as segment A.

As a second step the four criteria are integrated into one value of association strength. To do that a supervised learning approach is used: Training data is generated manually by taking human ordered segments. The task of association is then modeled as a binary classification problem and an SVM classifier is trained to solve this problem.

Finally a bottom-up approach is used to order the sentences. The strategy is basically a greedy strategy: The pair of sentences (or segments) with the highest association strength is chosen and included in the new ordering. Afterwards the next pair gets chosen, until all sentences/segments are ordered.

Aspect we liked most Obviously some of the criteria are very similar to strategies we've seen in other papers on the topic. Using a Machine Learning approach to combine them is a very interesting idea though, as it is very hard to figure out which ordering strategy might work best for the currently observed pair of sentences. A classifier can greatly help doing that.

Possible improvement The final step is in essence just a greedy approach, choosing the sentences, that currently have the highest association. This might be improved by other strategies. For example an approach similar to that of the augmented algorithm in (Barzilay et al., 2002) might help by clustering the sentences into blocks, which are ordered separately.

3.2 Cluster Adjacency

The second paper we discuss is "Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization" (Donghong and Yu, 2008).

Problem definition The problem addressed is basically the same as in our project: After selecting sentences for a summary it is necessary to find an ordering of those sentences such that the resulting summary is well readable.

Main approaches After choosing which sentences should be in the summary the first step is to cluster all the sentences in the source documents (not only the chosen ones) into K clusters, where K is the number of chosen sentences.

After that it is possible to calculate the adjacency of each pair of sentences by calculating the adjacency of the clusters the sentences belong to. The ordering is then achieved by taking a greedy approach on the adjacency: Assuming the sentences $S_1, \ldots, S_i$ are already ordered, then S_{i+1} would be the sentence most adjacent to S_i.

Aspect we liked most A nice aspect about this paper is, that no additional parameters are necessary. Given a set of sentences and some way of choosing the ones, that should appear in the summary the algorithm does not need any further "magic" input.

Furthermore the evaluation method used does not only rely on asking human judges about the quality of the summary, but instead a quantitative approach is found by comparing how far the generated ordering is away from the human ordered reference. This is not an optimal evaluation as well, as there might be multiple equally good orderings, but it gives an idea and is comparatively easy to do, whereas asking judges is quite elaborate.

Possible improvement Like before the final ordering is done quite greedily. It might be possible to improve that by calculating some form of global optimum on the adjacency, and thus getting an order, that maximizes pairwise adjacency. However, this might well be computationally too expensive to be feasible.

3.3 Another Machine Learning approach

The last paper discussed is "A Machine Learning Approach to Sentence Ordering for Multidocument Summarization and Its Evaluation" (Bollegala et al., 2005).

Problem definition The problem addressed is once again sentence ordering for multi-document summarization. The paper is quite similar to the aforementioned (Bollegala et al., 2010), as it too uses a Machine Learning approach to find a measure of adjacency for sentence pairs.

Main approaches The first part of the algorithm is basically the same as the one in (Bollegala et

al., 2010). Again a Machine Learning method is applied to combine different criteria (called "experts") into one measure of adjacency. This approach uses one criterion more than the other one though: A probabilistic expert provides a probabilistic model which can be used to calculate the probability of one sentence appearing after another, thus implying an ordering. The classifier learned is used to decide, which ordering method to use for a pair of sentences, thus yielding a total preference function. However, finding an optimal order for this total preference function would be NP-complete, therefore a greedy approximation proposed by (Cohen et al., 1999) is used to order the sentences.

All three discussed papers have in common, that they use a quantitative approach to evaluate in addition to asking human judges for a rating.

Aspect we liked most Like before the combination of several ordering strategies is a very interesting idea. One thing that differentiates this paper from (Bollegala et al., 2010) is the use of another expert/criterion (the probabilistic one), that has been dropped in the newer paper. It would be interesting to know, why this was done.

Possible improvement The authors did not use a dependence structure in the algorithm, which two of the experts highly rely on. This might have been done to push the new "succedent" expert that does not need this information. However this might make the evaluation a bit biased towards one of the ordering methods and the algorithm might be improved by fixing this. Furthermore it would have been interesting, if the approach works on other domains than news as well. However news summarization is the only domain evaluated in the paper.

4 Conclusion

This report included two parts: We first detailed the current position of our project and then discussed and compared three related papers. All of the related papers discussed use an approach, that is quite different from the one considered for our project. Especially the Machine Learning approaches in (Bollegala et al., 2005; Bollegala et al., 2010) are quite deviating. The cluster adjacency method from (Donghong and Yu, 2008) however has some similarities: Both use some sort of clustering to do the ordering. The main difference here is, that the method in (Donghong and Yu, 2008) uses the clusters to order the already chosen sentences, whereas (Barzilay et al., 2002) orders the clusters and the sentences inside of the clusters to achieve an ordering on sentences which are then chosen out of the clusters.

References

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35 – 55.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2005. A machine learning approach to sentence ordering for multi-document summarization and its evaluation. In *In Proceedings of IJCNLP*.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, 46(1):89 – 109.

W. W. Cohen, R. E. Schapire, and Y. Singer. 1999. Learning to order things. *Journal of Artificial Intelligence Research*, 10.

Ji Donghong and Nie Yu. 2008. Sentence ordering based on cluster adjacency in multi-document summarization. *Third International Joint Conference on Natural Language Processing*.