Reflections

Feelings for Writing a Manual Summary

For the multi-document summarization assignment, I spent different periods of time to the three steps in the instructions. I think all of the three steps are critical components of the workflow for both manual and automatic summarization task.

For step 1, it took me more time than the other steps because I not only selected the sentences that are related to the topic and query but also wrote down some key words as notes during the selection procedure. The outcome of step 1 is the premise for further advanced selection and manipulation. Due to its importance, I made a careful observation and analysis to the semantic expressions and structure of the sentences while I was doing the selection work. The total number of sentences of 25 stories is very large for humans to look through and manually make selections, but it is inevitable to achieve the summarization goal. After step 1, I selected 163 relevant sentences to the topic and query.

For step 2, the main duty is to identify the repetition of sentences based on the meanings and interrelationships. I did not spent much time on this step compared to others because I referred to the key words that already been written down during step 1 to select the sentences I need, and tried to pick sentences that best cover all the key words. In addition, compared to sentences within specific locations or examples, I chose the general sentences which best explain the causality logically, which is the preliminary step of summarization. Since the subset must not exceed 250 words, the most challenging one in this step is how to deal with the semantic overlapping issue and how to make a comprehensive summary as accurate as possible. After step 2, I selected 7 sentences as the extractive summary.

For step 3, I also mainly referred to the key words formerly written down. What I thought about most were the relationship of those key words such as the classification of different aspects and the structure for the composition of the summary. I divided the impacts of global climate change into four main aspects: physical variation, ecological system deterioration, human health and society development. Then, I added more details to the four aspects and completed the summary eventually. I think the work in this step mainly relies on human's thinking and organization of languages.

Thoughts for Automatic Summarization

The main part for automatic summarization is the selection of sentences. After completing the summary according to the required workflow, I have some ideas for the computer system to accomplish this task automatically.

Firstly, computer could select the sentences that match appropriately to some key words. There are some theme related words such as 'global climate change', 'impact' and 'implication'. Verbs that reflect the variation could also be targeted words, such as 'decrease' and 'increase', which implies the impacts of variation in quantity. Words and phrases that convey the reason and result relationships should also be included, such as 'result from/in', 'play a role', 'exacerbate', 'emerging', 'caused by', 'as a result', 'lead to' and 'affect'. Those key words could help to identify the sentences related to certain topic and query, and also help to find the causality within the sentences.

Secondly, it is a good idea to reduce redundancy and repetition though the combination of synonyms. The list of synonyms may be discrepant among different topics. And the creation of the list needs the help of linguistics. As with most linguistically motivated systems, there is a dependency grammar representation of each sentence (Blake, 2007). For this task, 'global climate change' and 'global warming' could be

regarded as synonym logically. Also, it is ideal that the system could automatically recognize words that could be identified as one concept after reading all the sentences in the pool. The automatic combination of relevant sentences with the same meaning and selecting the most appropriate one are another essential facets to be considered in the system.

As Blake et al. demonstrated, document pre-processing is the first step before sentence selection. Systems are devised to use a variety of tools to generate a dependency grammar, such as Collins Parser and Stanford Parser. For Stanford NLP, it can divide the whole text into individual sentences and identify the person, location, organization and other roles in the sentences, which could make contributions to further lexical analysis (2007).

After pre-processing, query expansion (QE) could be a powerful way to improve the recall of relevant sentences, but often will lead to decreased precision (Blake, 2007). Query expansion based only on nouns is a relatively easy way. The first step could use a lexicon to identify the original and base form of each topic and query term; for example, in this summarization task, the term ‘impacts’ can be expanded to ‘impact’. Apart from nouns, the transformation in grammar for verbs could also be stemmed; for example, ‘are emerging’ can be expanded to the term ‘emerge’. In the second step, the system could also remove stop words such as ‘the’, ‘a’, ‘and’ from the pool of terms.

For multi-document summarization task, lexical simplification can aid in summarization by removing sections of a sentence that do not contain essential information (Blake, 2007). In this section, linguistics would be the main reference for manipulation. The system could remove noun appositive, gerundive clause, non-restrictive relative clause, intra-sentential attribution, lead adverbials and remove extra words, adverbs, attributable information, joining words and gerund phrases (Blake, 2007). As for relative clauses like ‘whom’, ‘which’ and ‘when’, it is good for the system to be able to identify its relationships.

There should also be some advanced resolution and analysis inside one sentence. To identify syntactically valid sub-sentences, the system could extract all branches that contain a subject and an object, for instance, branches of complete sentences concatenated by ‘and’ or ‘but’. Besides, adverbial phrases, such as ‘Also’ and ‘In fact,’ are often used to open a sentence, but typically do not provide important information. The system should identify those phrases deactivate its meaning. Among the sentences in the document tab of the spreadsheet, I find some cited sources of quoted material such as ‘…said’, but manual summaries rarely use this information. So the system should remove this part from the original sentence.

At last, a variety of feature selection methods are used to select the identical and meaning words for summarization. To rank these words, some weighting schema can be adopted such as tf*idf method. Making sentences to develop a fluent summary based on these selected terms are the last step which involves sophisticated linguistics knowledge to achieve the task.

References

Blake, C., Kampov,J., Orphanides,A., West,D., & Lown,C. (2007) UNC-CH at DUC 2007: Query Expansion, Lexical Simplification, and Sentence Selection strategies for Multi-Document Summarization, Presentation at Document Understanding Conference (DUC) 2007, Rochester, NY.