한국컴퓨터정보학회 하계학술대회 논문집 제23권 제2호 (2015. 7)

어휘의 동시 발생 빈도와 분포를 이용한 다중 주제 회의록 요약

이병수°, 이지형*
°성균관대학교 정보통신대학 DMC공학과, °삼성전자, *성균관대학교 정보통신대학 전자전기컴퓨터공학과
e-mail : {bs.lee, john}@skku.edu°*, bs0425.lee@samsung.com*

Multi-Topic Meeting Summarization using Lexical Co-occurrence Frequency and Distribution

Byung-Soo Lee°, Jee-Hyong Lee*
°Dept. of DMC Engineering, Sungkyunkwan University, °Samsung Electronics,
*Dept. of Electrical and Computer Engineering, Sungkyunkwan University

● Abstract ●

본 논문에서는 어휘의 동시 발생 (co-occurrence) 빈도와 분포를 이용한 회의록 요약방법을 제안한다. 회의록은 일반 문서와 달리 문서에 여러 세부적인 주제들이 나타나며, 잘못된 형식의 문장, 불필요한 잡담들을 포함하고 있기 때문에 이러한 특징들이 문서요약 과정에서 고려되어야 한다. 기존의 일반적인 문서요약 방법은 하나의 주제를 기반으로 문서 전체에서 가장 중요한 문장으로 요약하기 때문에 다중 주제 회의록 요약에는 적합하지 않다. 제안한 방법은 먼저 어휘의 동시 발생 (co-occurrence) 빈도를 이용하여 회의록 분할 (segmentation) 과정을 수행한다. 다음으로 주제의 구분에 따라 분할된 각 영역 (block)의 중요 단어 집합 생성, 중요 문장 추출 과정을 통해 회의록의 중요 문장들을 선별한다. 마지막으로 추출된 중요 문장들의 위치, 종속 관계를 고려하여 최종적으로 회의록을 요약한다. AMI meeting corpus를 대상으로 실험한 결과, 제안한 방법이 baseline 요약 방법들보다 요약 비율에 따른 평가 및 요약문의 세부 주제별 평가에서 우수한 요약 성능을 보임을 확인하였다.

키워드: 회의록 요약(Meeting Summarization), 문장 추출(Sentence Extraction), 중요 단어 생성(Keyword Generation)

I. Introduction

문서요약이란 문서가 담고 있는 핵심 의미를 유지하면서 문서의 길이를 줄이는 작업으로 크게 추출요약과 생성요약으로 구분할 수 있다. 추출요약은 문서에서 중요하다고 생각하는 문장만을 선별하여 요약문을 제공하는 것이다. 이에 비해 생성요약은 문서에서 중요하다고 생각하는 일부분을 선택한 후, 자연어 처리 기법을 이용하여 새로운 문장으로 요약문을 제공하는 것이다. 그러나 현재의 자연어 처리 기술 한계와 제한점 때문에, 상대적으로 접근과 구현이 쉬운 추출요약에 관한 연구가 주로 이루어지고 있다[1].

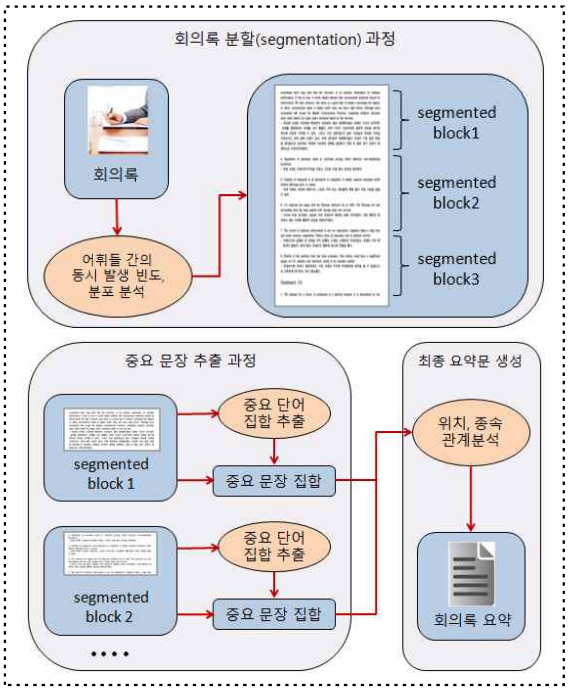
추출 요약은 뉴스 기사나 논문 등 잘 구성된 텍스트 도메인에 대해서는 좋은 성능을 보이고 있으나, 회의나 전화 대화 같은 직접적인 음성 도메인을 기록한 문서는 잘못된 형식의 문장, 불필요한 잡담들을 포함하고 있으며, 문서에 여러 세부가 나타날 수 있기 때문에 이러한 특징들이 문서 요약과정에서 고려되어야 한다.

기존 회의록 요약방법으로 문서에서 선별한 중요 문구와 높은 순위로 연관된 문장들을 나열함으로써 요약문을 추출하는 방법이 있다[2]. 단어의 빈도수만을 고려하여 중요 단어를 선택하는 방법에 비해 성능 향상이 있었으나, 문서에 여러 세부적인 주제가 있을 경우 이를 모두 포함하지 못하는 문제가 있어 이 부분에 대한 보완이 필요하다. N. Garg와 B. Favre는 그래프 기반의 TextRank[3] 알고리즘을 확장하여 문장들을 클러스터로 만들고 클러스터 사이의 그래프를 구성하여 문서를 요약하는 방법을 제안하였다[4]. 기존 TextRank 방법에 비해 좋은 성능을 보이지만, 문장 개수만큼의 많은 클러스터가 존재하는 경우 TextRank 방법과 유사해지는 한계점도 있다.

본 논문에서는 어휘의 동시 발생 (co-occurrence) 빈도와 분포를 이용하여 앞서 살펴본 회의록 특징들을 고려한 요약 방법을 제안한다.

II. Legislation Case of Marital Property Regime

본 논문에서 제안하는 회의록 요약 모델의 각 단계별 수행절차는 [그림 1]과 같다.

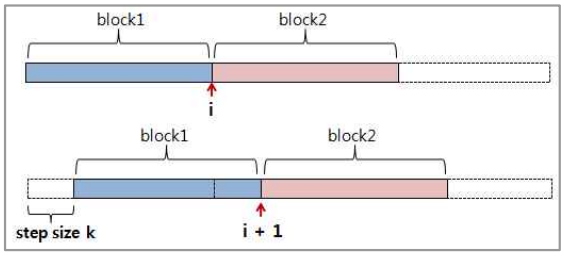


The meeting summarization model

먼저 여러 세부 주제가 포함된 회의록의 특징을 고려하여 회의록의 각 어휘들 간의 동시 발생 (Co-occurrence) 빈도를 이용하여 회의록 분할(segmentation) 과정을 수행한다. 다음으로 각 분할된 영역 (block) 단위로 중요 단어 집합을 생성하며, 생성된 중요 단어 집합과 문장 간의 유사도를 계산하여 유사성이 높은 문장 순으로 각 분할된 영역 (block)의 중요 문장을 추출한다. 마지막으로 선택된 문장들의 위치 및 종속 관계를 고려하여 최종적으로 회의록을 요약한다.

1. 회의록 분할 (segmentation) 과정

회의록 분할 (segmentation) 과정의 전처리 단계로 문서에서 불용어 (stopwords)를 제거하고 Porter's stemming algorithm[5]을 적용한다. 분할 (segmentation)의 목적은 각 분할된 영역 (block)이 의미 있는 단위로 동일한 주제를 갖도록 분할하는 것으로 문장 구별 없이 Text를 token sequence로 분할한 후, 일정한 수 (window size)의 token을 가진 sequence를 하나의 block으로 간주한다[6]. [그림 2]와 같이 block1, block2는 일정한 step size k에 해당하는 token의 개수만큼 순차적으로 이전 token을 삭제하고 다음 token을 추가하면서 어휘 유사도를 바탕으로 각 token 지점 (i, i+1, i+2, i+3, ...)에 대한 점수를 계산한다. 최종적으로 threshold bottom 이하로 점수가 부여된 token 지점을 후보군으로 하여 미리 정의된 neighbors 구간 안에 더 작은 유사도 점수가 부여된 token 지점이 없다면, 해당 지점을 topic boundary로 선택한다.



token sequence

본 논문에서는 training set을 통해 window size와 step size, neighbors는 각각 150, 5, 60으로 설정하였다.

2. 중요 단어 집합 추출 과정

회의록과 같은 대화 언어에서 핵심 단어가 반복되는 특징을 고려하여, 단어의 빈도수와 POS tagger를 활용한 방법[2]을 사용한다. 대화록에서 사용된 불필요한 단어 (e.g. "hm", "gonna", "ooops")와 불용어 (stopwords)를 제외한 후, 형용사와 명사의 정규 표현식과 일치하는 n-gram을 (n=1,2,3) 추출한다.

$$weight = (n-gram\text{의 빈도수}) \times n \quad (1)$$

추출한 $n-gram$ 길이에 가중치를 주게 되며, 최종적으로 식 (1)을 이용해 점수를 부여한다. 식 (1)의$weight$ 은 해당 $n-gram$의 최종 점수이고, n 은 $n-gram$ 길이이다.

3. 중요 문장 추출 과정

중요 단어 집합과의 유사도 점수가 높은 문장 순으로 중요 문장을 추출하게 되며, Maximum Marginal Relevance (MMR)[7] 방식을 사용한다. 문장과 중요 단어 집합과의 유사도 점수 산정 방식은 식 (2)와 같으며, 이미 중요 문장으로 선택된 문장들과 중복된 문장이 선택되는 것을 방지하기 위해 balance factor λ를 사용한다.

$$score = \lambda \times sim_1 - (1-\lambda) \times sim_2 \quad (2)$$

$$sim_1(u) = \frac{(\sum_i occ(g_i,u) \times w_i)}{\log(1+length(u))} \quad (3)$$

$$sim_2(u,V) = \frac{|u \cap V|}{\max(|u|,|V|)} \quad (4)$$

식 (3)에서 $occ(g_i,u)$는 문장 u에 포함된 $n-gram$ g_i의 빈도수이며, w_i는 해당 $n-gram$의 $weight$ 값이다. 문장 길이에 대한 log 값을 적용해 상대적으로 더 짧은 문장에 가중치를 주도록 한다. 식 (4)의 $sim_2(u,V)$는 문장 u와 이미 선택된 문장 V 사이의 유사도 점수를 산출하며, 공통된 단어를 기준으로 더 긴 문장의 길이로 normalize 한 값을 사용한다.

4. 최종 요약문 생성

각 분할된 영역 (block)에서 추출한 문장들의 영역 (block) 위치

및 종속 관계를 고려하여 최종적으로 회의록을 요약한다. 한정된 요약문에 가능한 많은 정보를 포함할 수 있도록 각 영역 (block)에서 추출하는 요약문 길이는 해당 영역 (block)의 길이에 비례하도록 제한하는 proposed1 (block length based ratio) 방법과 분할된 영역 (block)들의 평균 centroid vector를 설정 후 centroid vector와 해당 영역 (block)의 코사인 유사도 (cosine similarity) 점수에 비례하여 영역 (block)의 추출 문장 길이를 한정하는 proposed2 (block similarity based ratio) 방법을 사용한다.

III. 실험 및 평가

1. 실험 데이터 및 성능 평가 함수

실험 데이터는 AMI meeting corpus[8]를 사용하였으며, 20개의 회의록 (ES2004, ES2014, IS1009, TS3004, TS3007)을 test set으로 96 개의 회의록을 training set으로 사용하였다. 각 회의록에는 전문가가 작성한 하나의 요약문이 있으며 평균 요약문의 길이는 해당 회의록의 약 6%이다.

성능 평가를 위해 두 가지 baseline과 비교하였다. baseline1은 요약문의 제한 길이가 만족될 때까지 각 단계에서 가장 긴 문장을 선택해 나가는 요약 방법 (longest sentence)이고, baseline2는 문서 전체의 어휘 발생빈도를 기반으로 중심 어휘 벡터를 설정한 후, 중심 어휘 벡터와 각 문장과의 코사인 유사도 (cosine similarity) 점수를 바탕으로 MMR을 사용하여 중심 문장을 추출하는 방법 (centroid based MMR)이다[9]. baseline2의 MMR balance factor λ 는 training set 실험을 통해 가장 성능이 좋게 나타난 0.3으로 설정 후 test set 실험을 진행하였다.

요약문의 전체적인 성능은 ROUGE-1[10] 점수로 평가하였으며, F_1-measure는 식 (5)와 같이 정확률 (precision)과 재현율 (recall)을 하나의 값으로 표현한다. 세부주제를 고려한 평가 방법으로는 전문가가 작성한 요약문을 주제별로 분리하여 각각의 주제에 대한 재현율 (recall) 점수로 성능을 평가하였다.

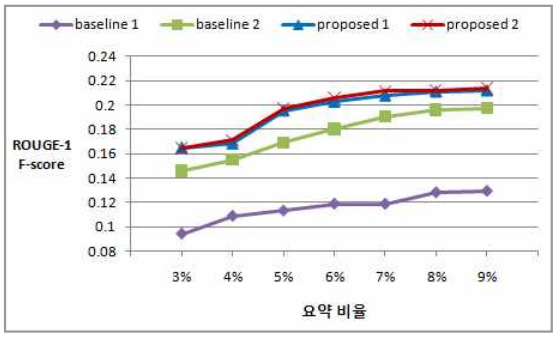
$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

2. 실험 및 결과

각 회의록에서 3개-5개의 분할된 영역 (block)을 생성하였고, 각 영역 (block)에 대해 중요 단어 집합을 상위30개씩 추출 하였다. 각 분할된 영역 (block)에서 추출할 수 있는 중요 문장의 길이는 proposed1과 proposed2 방법을 적용하여 두 가지 방법에 대한 성능을 각각 측정하였으며, proposed1, proposed2 모두 training set 실험에서 가장 좋은 성능을 보인 0.1로 MMR balance factor λ를 설정하였다.

2.1 요약 비율(%)에 따른 평가

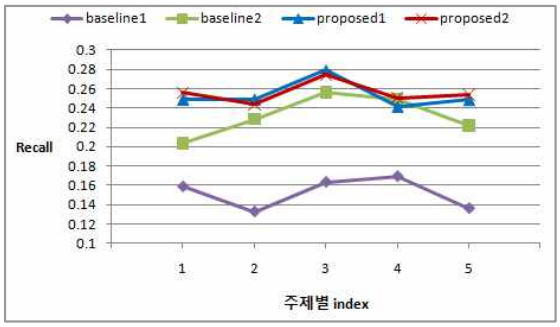
먼저 요약 비율을 3%-9% 구간에서 1% 구간별로 요약문을 생성하고 그 결과를 분석하였으며, 정확한 평가를 위해 불용어 (stopwords)는 평가에서 제외하였다. [그림 3]은 전문가 요약문과 비교한 ROUGE-1 F-measure 평가 결과이다. baseline 방법들에 비해 proposed1과 proposed2 방법을 사용한 요약문이 요약 비율 전 구간에서 성능이 높게 나타났음을 볼 수 있다. 또한 요약 비율에 상관없이 proposed2의 요약 결과가 대체적으로 proposed1에 비해 높은 성능을 보인다. 이를 통해 분할된 영역 (block)의 길이에 비례해서 중요 문장 추출 길이를 정하는 것보다, 중요한 내용이 포함된 영역 (block)을 선별해서 중요도에 따라 영역 (block)별 추출할 문장 길이를 조절하는 것이 의미가 있음을 알 수 있다.



Performance variation with summary ratio

2.2 세부 주제별 평가 결과

전문가가 작성한 요약문을 주제별로 5개로 분리하였으며, [그림 4]는 각 주제에 대한 Recall (재현율) 측정 결과이다.



Recall result of each subject

[그림 4]에서 proposed1과 proposed2 방법을 사용한 요약문이 baseline 방법들 보다 각 주제별로 전 구간에서 고른 성능을 보이며, 성능이 높게 나타났음을 확인할 수 있다. baseline 방법들이 상대적으로 회의록의 처음과 마지막 부분의 주제를 잘 반영하지 못하는 반면에, proposed1과 proposed2는 각 주제별로 분할 (segmentation)을 이용해 회의록을 요약함으로써, 세부 주제를 잘 반영하여 요약한다는 것을 알 수 있다.

IV. Conclusion

본 논문에서는 회의록의 특징을 반영하여 어휘들 간의 동시 발생(co-occurrence) 빈도와 분포 정보를 바탕으로 회의록의 분할(segmentation)을 이용한 요약 방법을 제한하였다. 실험 결과, 요약 비율에 따른 평가 및 세부 주제를 고려한 평가에서 제안한 방법이 baseline 방법들에 비해 높은 성능을 보였으며, 회의록의 세부 주제 및 흐름을 파악하는데 도움을 준다는 것을 보였다.

제한된 회의록 요약문의 길이로 인해 분할된 영역 (block) 별로 추출할 수 있는 중요 문장의 길이를 정하는 것은 중요한 문제이다. 향후 연구로 각 영역 (block)에서 추출하는 문장의 길이를 한정하는 방법을 개선하여 추가적으로 성능을 높일 수 있는 연구를 진행할 예정이다.

References

[1] I. Mani, “Automatic summarization,” John Benjamins Pub Co., 2001.

[2] K. Riedhammer, B. Favre, and D. Hakkani-Tur, “A Keyphrase Based Approach to Interactive Meeting Summarization,” in Proc. IEEE Workshop on SLT, pp.153-156, Dec. 2008.

[3] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Texts,” in Proc. EMNLP, pp.404-411, 2004.

[4] N. Garg, B. Favre, and K. Riedhammer, “ClusterRank: A Graph Based Method for Meeting Summarization,” in Proc. InterSpeech, pp.1499-1502, 2009.

[5] M. Porter, “The Porter Stemming Algorithm,” http://www.tartarus.org/martin/PorterStemmer, 1980, [Accessed: January 10, 2015]

[6] M. A. Heart, “Multi-Paragraph Segmentation of Expository Text,” in Proc. ACL, pp.9-16, 1994.

[7] J. Carbonell and J. Goldstein, “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries,” in Proc. ACM SIGIR, pp.335-336, 1998.

[8] J. Carletta, S. Ashby, and S. Bourban et al., “The AMI Meeting Corpus: A Pre-Announcement,” in Proc. MLMI, pp.28-39, 2005.

[9] D. Gillick, K. Riedhammer, B. Favre , “A Global Optimization Framework for Meeting Summarization,” in Proc. IEEE ICASSP, pp.4769-4772, 2009.

[10] C. Lin, “Rouge: A Package for Automatic Evaluation of Summaries,” in Proc. ACL Text Summarization Workshop, pp.74-81, 2004.