질의응답서비스를 위한 복수 응답 요약에 관한 연구

A Study on Summarizing Multi-Answers for Question Answering Service

최상희, 연세대학교 문헌정보학과, tudultudul@yonsei.ac.kr

Choi, Sang-hee, Dept. of Lib. & Info. Sci., Graduate School of Yonsei University

이 연구에서는 질의응답을 기반으로 한 검색 서비스를 이용할 때 이용자가 효율적으로 응답정보를 이용할 수 있도록 검색되는 복수 응답을 요약하는 방안을 제시하였다. 복수 응답을 요약하기 위해서는 질의중심방식과 응답중심방식이 비교되었다. 생성된 요약문을 평가한 결과 응답내용을 중심으로 요약하는 방식이 질의중심으로 요약하는 방식보다 질의에 적합한 문장을 효과적으로 추출하고 중복되는 정보도 줄여주는 것으로 나타났다.

1. 서론

질의응답서비스는 질의를 가지고 있는 이용자와 그 질의에 대한 답을 알고 있는 이용자를 연결하여 정보를 제공하는 서비스로서, 이용자가 참여하여 정보를 생산해내는 이용자 제작 컨텐츠(User Created Contents)의 대표적인 유형이다. 특히, 최근 들어 포털서비스에서는 다양한 주제로 쉽게 질의를 게시할 수 있게 하고 많은 이용자들이 편리하게 답을 할 수 있도록 질의응답서비스를 활성화시키고 있다. 점차 질의응답서비스에 질의응답내용이 축적되면서 정보문제를 가지고 있는 이용자들은 질의응답서비스를 이용할 때 직접 질의를 게시하고 답변을 기다리기 보다는 이미 대용량으로 축적되어 있는 질의응답내용을 검색하여 필요한 정보를 알아내고 있다. 따라서 질의응답서비스에서 일반 웹문서 검색과 같이 기존에 축적되어 있는 질의응답내용 중에서 현재 이용자가 가진 질의와 가장 유사한 질의응답내용을 검색해주는 것이 중요하다. 그러나 질의응답서비스에 축적되어 있는 질의와 응답내용을 보면 대부분의 질의응답서비스에서 이용자가 같은 질의를 게시하거나 같은 답변을 기술할 때 제한을 하지 않기 때문에 같은 내용이 중복되어 있는 경우가 많다. 그 결과 질의응답서비스에서 축적되어 있는 기존 질의응답내용을 검색했을 경우, 유사한 내용의 질의응답정보가 많이 검색되기 때문에 이용자는 일일이 내용을 확인한 후 찾고자 했던 응답내용을 다시 찾아내야 한다. 이와 같이 새 질의를 게시하지 않고 기존에 처리된 질의응답내용을 검색할 때에는 일반적인 웹문서를 검색하는 것과 같이 대량으로 검색된 결과에서 적절한 정보를 찾아내는데 문제가 발생하고 있다.

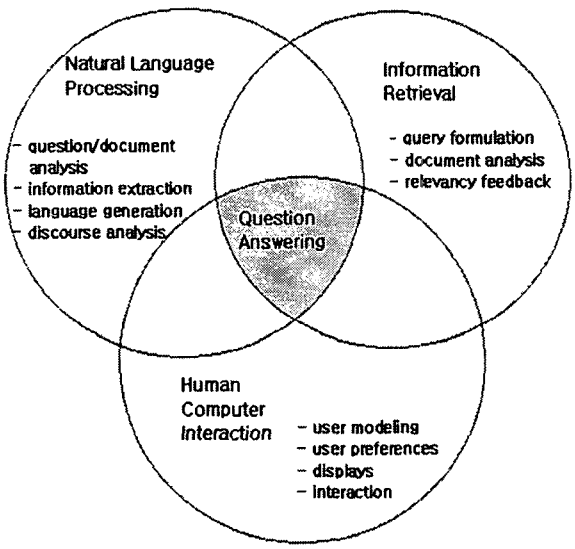
이 연구에서는 질의응답서비스에 축적되어 있는 질의응답내용을 검색하여 활용할 때 이용자가 검색된 많은 응답내용을 비교하는데 들이는 시간과 노력을 최소화하고자 검색된 여러 응답내용을 요

약하여 제공하는 방법을 제시하고자 한다. 응답내용을 요약하는 과정은 일반적인 복수문서 요약 방식 중 단락을 추출하여 요약을 생성하는 방식을 기반으로 수행되었고 단락을 추출하는 과정은 질의중심과 응답중심 등 두 가지 방식으로 적용하여 각 방식의 효율성을 평가하였다.

2. 질의응답과 자동문서요약

질의에 대한 응답을 자동으로 추출해주는 질의응답시스템이란 이용자들이 자연어로 표현된 질의어를 처리하는 인간 컴퓨터 상호과정에 기반한 것이다.

<그림 1>에서 나타듯이 질의응답과정은 자연어처리과정과 정보검색, HCI(Human Computer Interface) 등 다양한 정보처리기법과 기술이 결합되어 완성되는 과정이라고 할 수 있다(Maybury 2004).



<그림 1> 질의응답과 정보검색 분야간 관계

질의에 맞는 응답을 추출하는 과정은 답이 있을 만한 적절한 문서나 데이터를 선별하여 검색하는 처리과정, 적합한 정보원에서 가능한 답을 추출한 후 검증하며 우선순위를 주는 처리과정, 마지막으로 효과적인 방법으로 응답을 표현하고 설명해주는 과정으로 구성된다.

자동문서요약은 이용자가 원하는 정보원에서 이용자가 찾고자 하고 중요한 정보를 축약된 형태로 제공하는 방법이다. 그 중 복수문서 요약은 기존에 분류된 문서집단이나(Radeve, Jing, and Stys, 2003) 검색 결과에서 포함된 문서들의 주제를 요약하는데 활용되는 요약을 생성하는데 적용되어 왔으나 최근 이용자 중심으로 요약을 생성하는 다양한 연구가 수행되었다(Maybury 2004). 문서요약에 이용자가 개입하게 되면 이용자중심요약(user-focused summarization) 또는 질의중심 요약(query-focused summarization)이라고 할 수 있다. 즉, 이용자 중심의 요약은 이용자가 알고 싶은 주제나 정보수준에 따라 요약을 생성해 주는 것이다(Tombros, and Sanderson 1998). 이용자 중심의 요약은 질의응답 과정과 유사한 것으로 답을 구하는 질의가 요약을 할 때 기준이 되는 이용자의 정보요구에 해당하는 것이며, 이용자의 주제, 즉 이용자 질의에 따라 요약된 내용은 응답내용으로 활용될 수 있는 것이다.

자동요약에서 대상이 되는 문서 내에서 정보단위는 용어, 문장, 단락 등이 있다. 이 중 단락(passage)은 자동으로 분할될 수 있는 일정량의 연결된 정보로서 문헌으로부터 추출한 연속된 텍스트이다(Kaszkiel, and Zobel 2001). 일반적으로 단락은 일련의 문장을 나열한 것으로 단순히 단답성의 정보외에도 정보의 맥락을 제공하고 있다. 단락에는 생성시기에 따라 정적단락과 동적단락으로 나뉠 수 있는데 정적단락은 주로 문서 내 주제에 따라 미리 분할되어 있는 단락을 말한다. 반면 동적단락은 단락을 발생시킬 기준, 즉 단락주제를 동적으로 이용자의 질의에 따라 규정한 다음 그 기준에 가장 적합한 정보를 문서에서 검색한 후 그 검색 결과를 중심으로 정보를 확장하여 단락을 추출하는 것이다. 질의중심의 요약을 하게 될 경우, 정적단락은 질의가 미리 분할해 놓은 주제단락에 대응되지 않을 수 있으므로 질의중심의 요약에 적

용되기에는 적합하지 않다고 할 수 있고 동적단락은 이용자 질의처럼 동적으로 변화하는 특성을 가진 요소에 적합하다.

동적단락을 추출하기 위해 활용될 수 있는 정보검색기법으로는 개념확장 활성화 기법이 있다. 개념확장 활성화 기법은 인공지능기반 시스템에서 초기노드와 연결된 노드를 따라 항해하면서 적합한 정보를 검색해나가는 기법이다. 개념확장 활성화 기법 중 비교적 좋은 성능을 나타내고 있는 bnb (branch-and-bound search)기법은 개념확장이 진행되는 동안 최단 경로를 찾기 위한 방법이다(김정하 2002). bnb 확장 활성화 기법은 확장 중심노드에서 가장 유사한 방향으로 노드를 추적해 가는 방식이므로, 특정 문장을 중심으로 가장 유사도가 높은 문장으로 확장해단락을 추출할 수 있다. 질의응답에 적용될 때는 질의에 가장 적합한 문장을 중심으로 주변 문장으로 확장시키는 과정을 통해 복수 응답내 요약문에 포함될 만한 주요 문장을 탐색하여 최종 요약을 생성하는 것이다. 생성된 최종요약은 질의에 적합한 응답내용에 해당한다.

3. 복수 응답 요약 실험 개요

복수 응답 요약 실험과정은 질의중심 요약과 응답중심 요약 두 가지 방식으로 수행되었고 실험을 위해 질의집단과 응담내용집단은 다음과 같이 구축되었다.

요약의 기준이 되는 질의집단은 포털서비스의 질의응답서비스 네이버 지식iN에 11개 주제별로 게시된 질의 중 무작위로 3개씩 추출하여 구축하였다. 질의집단으로 구축된 질의는 총 33개이다.

요약을 하는 응답내용집단은 질의집단으로 구축된 각 질의를 가지고 검색한 결과에서 상위 30개를 추출하여 총 330건의 문서집단이 구축되었다.

요약을 생성하는데 적용된 기법은 개념확장 활성화 방식인 bnb (branch-and-bound search)기법을 응용한 순차적 단락확장 기법이다. 순차적 단락확장 기법은 질의 중심으로 요약을 할 때 요약 정확률도 유지하면서 질의에 적합한 정보를 상대적으로 적게 중복되게 요약문에 포함시키는 것으로 평가되었다(최상희 2004). 순차적 단락확장 기법은 요약의 주제에 해당하는 문장을 단락확장 기준이 되는 중심문장으로 규정하고 주변문장에서 최단거리에 있는 문장을 하나씩 단락으로 포함시켜 추출하는 방식이다. 이때 포함된 최단경로에 있는 문장은 다음 확장에 포함되며 2차 확장을 위해 다른 경로와 비교할 때 다시 최단경로에 있는 주변문장를 선택하는 순환과정을 통해서 단락확장이 이루어지는 방식이다. 단락확장은 문장단위 유사도 비교를 기반으로 하였고 적용된 유사도 계수는 내적계수이다. 문장단위 유사도 산출을 위해 적용된 용어가중치는 역문장 빈도이다.

$$\text{내적유사도}(x, y) = \sum_{i=1}^{t} x_i \cdot y_i$$

$$\text{역문장빈도}(isf) = 1 + log 2 \frac{NS}{sf}$$

NS : 문장 집단 내 문장의 총수

sf : 문장빈도

생성되는 요약문은 800자를 기준으로 하였다, 이는 실험집단으로 구축된 응답내용이 평균 818자로 나타났기 때문이다.

3.1 질의중심요약

질의중심 요약 생성은 다음과 같은 과정으로 진행되었다.

① 질의와 비교하여 유사한 실험집단의 응답내용 내 문장을 검색, 추출한다.

② 검색된 문장을 유사도 순으로 배열하여 상위 3개의 문장을 1차 확장 중심문장으로 선정한다.

③ 중심문장을 기본으로 앞의 두 문장을 머리단락, 뒤의 두 문장을 꼬리단락으로 규정한다.

④ 중심문장을 기준으로 머리단락과 꼬리단락의 유사도를 측정하여 유사도가 높은 쪽의 단락을 대기열에 포함시킨다.

⑤ 대기열을 유사도의 내림순으로 정렬한다.

⑥ 중심문장을 기본으로 앞의 두 문장을 머리단락, 뒤의 두 문장을 꼬리단락으로 규정한다.

⑦ 중심문장을 중심으로 머리, 꼬리단락을 비교하여 유사도가 높은 단락을 요약으로 추가한다.

⑧ 추가된 단락의 문장은 중심문장으로 추가된다.

⑨ 요약문에 해당하는 수의 문장이 추출될 때까지 반복하여 수행한다.

3.2 응답중심요약

응답내용은 이용자가 질문한 질의에 대하여 다른 이용자가 답한 것을 축적한 것이다. 이용자가 질의를 한다는 것은 자신이 물어보는 내용에 대해 정확히 알고 있지 못하기 때문에 문의를 하는 것이므로 질의 주제를 효과적으로 표현하지 못하는 것으로 나타났고, 반면 응답자는 질의에 대한 답을 할 정도의 지식을 가지고 있기 때문에 오히려 질의주제에 적합한 내용을 기술하는 것으로 나타났다(최상희 2005). 이와 같이 질의응답과정에서 이용자의 지식부족으로 질의가 적절하게 기술하지 못했을 가능성이 있기 때문에 질의를 중심으로 요약을 하게 되었을 경우 효과적이지 못할 수 있다. 이 연구에서는 질의에 대한 보정방안으로 응답내용을 활용하여 요약을 하는 방안을 비교하고자 한다. 즉, 처음 질의를 한 사람이 생각하지는 못하여 질의어에는 포함되어 있지 않았지만 응답자가 질의자의 의도를 파악하고 질의 주제를 적절한 용어로 다시 표현하였을 가능성이 있기 때문에 응답내용이 오히려 요약을 하는 기준으로 활용되는 것이 바람직할 수 있다.

응답중심요약 생성은 다음과 같은 과정으로 진행되었다.

① 질의에 대해 검색된 상위 30개 응답내용에 포함된 문장을 비교하여 센트로이드를 생성한다.

② 생성된 센트로이드와 응답내용의 문장간 유사도를 비교하여 유사도 순으로 배열한 다음 상위 3개의 문장을 1차 확장 중심문장으로 선정한다.

③ 이하 질의중심 요약의 ③ - ⑨ 단계를 적용한다.

4. 실험결과 및 분석

요약문의 품질을 평가하는 평가척도로 적용한 평가척도는 요약정확률과 요약문장 중복률이다. 요약정확률은 생성된 요약문내 적합한 정보가 포함되어 있는 비율을 측정한 것이고 요약문장 중복률은 중복되는 문장이 포함되어 있는 비율을 측정한 것이다. 요약 정확률은 생성된 요약문내에 포함되어 있는 문장을 질의와 비교하여 질의에 적합한 문장 수를 측정하는 것이다. 즉, 질의에 답이 되는 정보가 포함되어 있는 정도를 평가하는 것이다.

요약문장 중복률은 생성된 요약문내에 중복된 정보를 포함한 문장의 비율을 측정하는 것이다. 자동으로 생성된 요약문이 질의주제에 대한 적절한 정보를 포함하고 있더라도 같은 내용을 중복해서 다루고 있다면 이용자에게는 실질적으로 도움을 주지는 못하기 때문이다. 그러므로 적합성 평가만으로는 생성된 요약문의 효율성을 평가하기 위해서는 제공되는 요약문 내 중복되는 정보비율을 파악하는 것이 중요하다.

$$\text{요약 정확률} = \frac{\text{질의에 적합한 문장 수}}{\text{요약문내 총 문장 수}}$$

$$\text{요약문장중복률} = \frac{\text{정보가 중복되는 문장 수}}{\text{요약문내 총 문장 수}}$$

< 표 1> 에서 나타났듯이 성능 평가 결과 응답내용 중심으로 요약문을 생성하는 방식이 질의를 이용하는 방식보다 질의에 적합한 문장을 효과적으로 추출하였다. 응답내용 중심으로 추출한 요약문에 적합한 문장이 포함되어 있는 비율은 0.732로 질의중심 방식의 0.687보다 높았다.

요약문내 문장들이 유사한 정보를 중복하여 제공하는 비율도 응답내용 중심 방식이 더 낮은 것으로 나타났다. 응답내용 중심의 중복률은 0.215이고 질의중심 요약방식은 0.271로 응답내용 중심 요약생성방식이 요약 정확률은 상대적으로 높게 유지하면서 낮은 중복률을 나타내고 있다. 질의를 중심으로 요약을 생성했을 때 적합정보를 추출할 때 같은 주제를 집중적으로 추출해낸다는 성향을 지니고 있다는 것이 판명되었다. 반면 응답내용 중심 방식은 정확률을 유지할 정도로 유사한 정보를 요약문에 포함시켜나가지만 상대적으로 다양한 정보로 확장해나가는 성향이 있다.

< 표 1> 복수 응답 요약 평가

	요약 정확률	요약문장 중복률
질의중심요약	0.687	0.271
응답중심요약	0.732	0.215

5. 결론

이 연구에서는 이용자가 질의에 대한 답을 효과적으로 이용하는 목적으로 복수 응답을 요약하는 방안을 제시하고 한다. 질의응답이 이루어지는 과정을 살펴보면 질의자는 지식부족으로 인해 질의주제를 질의로 효과적으로 표현하지 못할 수 있었고, 결과적으로 질의중심으로 요약을 했을 경우 적절하지 못한 정보가 포함될 수 있었다. 이에 대한 대응방안으로 이 연구에서는 응답내용을 활용하는 것을 고려하였다.

질의중심과 응답중심, 두 가지 방식으로 생성된 요약문의 성능을 평가한 결과 응답을 중심으로 요약문을 생성하는 방식이 질의중심 방식보다 질의에 적합한 문장을 효과적으로 추출하는 것으로 나타났다. 이는 질의를 중심으로 응답내용을 기반으로 요약하게 될 경우 적절하게 정보요구가 표현되지 못한 질의에 의존하게 되는 정도가 높아져 질의가 가지는 한계점을 드러나게 하는 것이다. 즉, 질의에 포함된 용어들이 질의 주제를 표현하는데 적절하지 못했을 경우, 그 용어들을 기준으로 요약을 생성했을 때 요약생성 성능에 영향을 미친다. 응답에서는 질의에서 적합하게 표현하지 못한 용어들을 응답자가 수정, 보완하여 기술할 수 있기 때문에 질의에서 성공적으로 표현되지 못한 질의 주제가 효율적으로 재표현될 수 있다. 그러므로 응답에 포함되어 있는 문장이 질의와 마찬가지로 요약을 생성하는 기준으로서 역할을 수행할 수 있는 것이다.

정보의 중복성 측면에서는 질의중심 방식이 적합정보를 추출할 때 같은 주제를 집중적으로 추출해낸다는 성향을 나타내었다. 반면 응답내용의 센트로이드를 이용하는 방식은 요약 정확률은 질의에 의존하는 방식과 유사하게 유지하면서 상대적으로 낮은 정보중복률을 제공하는 것으로 평가되었다. 응답중심방식은 질의중심방식보다 요약문을 추출하는 중심문장이 다양한 내용으로 확장되는 것으로 나타났다. 다양한 내용의 중심문장으로 주변문장을 확장, 추출하여 요약문에 포함시키게 되므로 요약문 내용에 중복되는 정보가 줄어들게 되었다.

이와 같이 질의응답서비스에서 질의부분에 의존을 하여 요약을 하게 되면 정확률도 저하되면서 다양성을 제공하지 못할 가능성이 있다. 따라서 질의응답서비스에서는 기존의 질의중심 요약 접근방식보다는 다양한 시각으로 응답내용을 활용하는 방안을 고려해야 할 것이다.

이밖에도 이 연구에서 도출된 결과는 다음과 같이 향후 연구를 할 때 고려할 수 있다

복수응답 요약에서 나타난 정보의 중복성은 응답내용의 신뢰성을 보완하는 방법으로도 적용될 수 있다. 즉, 중복도가 높은 정보는 그만큼 많은 이용자들이 질의에 대한 답변이라고 질의자에게 제시한 것이므로 가장 보편적인 답변으로서 신뢰도가 높아질 수 있다. 요약문에서 중복되는 정보를 여러 번 나열해서 보여주게 되면 이용자가 전체적으로 획득하는 정보량이 감소하게 되므로 문제가 발생할 수 있지만 중복되는 정보에 가중치를 주는 방법은 고려해볼만 하다. 또한 중복되는 정도를 다양하게 응용하여 정보 순위나 신뢰도 표시 등에 활용하는 것이 바람직 할 것이다.

참고문헌

김정하. 2001. 이용자 중심 요약문 생성에 관한 실험적 연구. 석사학위 논문, 연세대학교, 문헌정보학과.

최상희. 2004. 질의응답을 위한 복수문서 요약에 관한 실험적 연구. 박사학위. 연세대학교.

최상희. 2005. "문서구조를 이용한 질의응답문서 클러스터링에 관한 연구". 한국문헌정보학회지. 39(4):105-118

Kaszkiel, M., and Zobel, J. 2001. "Effective Ranking with Arbitrary Passages". *Journal of the American Society for Information Science and Technology,* 52(4): 344-364.

Maybury, Mark. T.2004. *New Directions in Question Answering*. Menlo Park: AAAI prss.

Radeve, D. R., Jing, Hongyan, Stys, Malgorzata, and Tam, Daniel. 2003. "Centroid-based Summarization of Multiple Documents"

<http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR 94-1438>

Tombros, Anastasios and Sanderson, Mark. 1998. "Advantages of Query Biased Summaries in Information Retrieval" In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* 2-10.