백과사전 질의응답 시스템을 위한 의미적 단락 생성 및 검색 기법

이충희 오효정 김현진 장명길
한국전자통신연구원 음성/언어정보연구부 지식마이닝연구팀
{forever, ohj, jini, mgjang}@etri.re.kr

Method of Semantic Passage Generation and Retrieval for Encyclopedia QA system

Chung-Hee Lee, Hyo-Jung Oh, Hyeon-Jin Kim, Myung-Gil Jang
Knowledge Mining Research Team
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문에서는 질의응답 시스템에서 질문의 주제와 개념적으로 일치하는 단락으로부터 정보를 추출할 경우에 보다 정확한 정답을 추출할 수 있다는 가정 하에 문장 주제를 활용한 의미적 단락 생성 및 검색 기법을 제안한다. 문장주제란 백과사전 문서 집합에서 공통으로 기술하는 내용이나 자주 언급하고 있는 사건 혹은 개념들의 집합을 의미하는 것으로, 주제별로 응집된 문장들로 재구성된 단락을 의미적 단락이라고 정의한다. 제안된 방법의 성능을 평가하기 위해 의미적 단락의 신뢰도를 파악하고, 백과사전 본문을 3문장 단위로 잘라서 고정길이 단락을 만든 후 의미적 단락의 검색결과와 비교하였다. 평가척도로는 TREC의 역순위평균(MRR : Mean Reciprocal Rank)과 상위 5개 단락 안에 정답유무를 측정하는 사용자 정답만족도를 사용하였다. ETRI 평가셋을 대상으로 한 실험 결과, 주제를 이용한 의미적 단락 검색 성능이 고정길이 단락 검색보다 우수함을 알 수 있었다.

1. 서 론

대부분의 질의응답 시스템에서는 정답 추출의 첫 번째 단계로 정답이 들어있는 문서를 검색해 주고, 다음 단계로 문서에서 정답을 추출해서 제시한다. 이때, 문서 단위가 아닌 단락 단위로 검색해서 정답을 찾을 수 있는데, 최근 연구에서 문서보다는 단락을 검색하여 정답을 찾는 것이 질의응답 정확도를 향상시킨다는 결과를 보여준다.[1] 그러므로 질의응답 시스템의 성능향상을 위해서 단락검색을 이용할 필요가 있다.

본 논문의 질의응답 대상인 백과사전은 인물, 정치, 예술 등의 특정 도메인 별로 표제어들이 분류되어 있고, 시간의 흐름에 의해서 본문이 구성되어 있다. 본문의 내용은 대부분 다른 주제를 가진 단문들로 구성되어 있으므로 본문 전체에 공통된 단어의 발생이 적고 인접한 문장들의 주제 응집력이 떨어진다. 그러므로 단락 분할 시에 공통 키워드 및 인접한 문장들의 주제 응집력을 이용하는 기존의 연구들을 백과사전 본문에 적용시키기에는 무리가 있다. 백과사전 본문 내용을 도메인 별로 분석해 보면 몇 가지 공통된 주제들이 발견되는데 이 주제를 본문의 각 문장들에 할당하고, 할당된 주제에 의해서 같은 주제의 문장들로 단락을 재구성하여 의미적 단락을 생성한 후에 단락검색에 이용할 경우, 질문의 주제와 동일한 주제의 단락에서 답을 찾을 수 있으므로 질의응답 시스템의 성능을 향상시킬 수 있다. 또한, 일부 주제(출생, 사망 등)의 경우, 다양한 형태로 문장 구성이 가능하므로 질문에 있는 키워드가 정답 문장에 한번도 안 나타나는 경우가 많으므로 키워드 매칭만으로는 정답 단락 및 정답을 찾을 수 없고 반드시 검색에 주제를 사용해야 한다.

이러한 질의응답 시스템 및 백과사전의 특성을 반영하기 위해서 주제에 의해서 재구성된 의미적 단락을 검색하는 방법을 본 논문에서는 제안한다. 의미적 단락검색의 성능을 평가하기 위해서 백과사전 본문을 일정 길이의 단락으로 나눠서 검색하는 고정길이 단락의 검색

결과와 비교하여 평가하고, 주제의 영향력을 평가하기 위해 주제를 이용한 경우와 이용하지 않은 경우를 비교 평가하였다. 마지막으로 각 주제별 정확도를 측정하여 정확도가 낮은 주제는 주제분류표에서 제거하여 재평가한 결과를 제시한다.

2. 관련연구

지금까지의 단락검색 연구에는 문서의 절, 단락 등의 형식적인 정보에 기반한 방법, 고정길이나 가변길이 윈도우에 의한 방법 그리고 의미적 실마리에 기반한 방법 등이 있다.

문서의 형식정보를 활용한 방법으로는 Salton[2] 등이 있는데, 문서의 길이가 매우 긴 경우에는 문서 검색보다는 문서의 절과 같은 형식적인 정보를 이용한 단락검색이 25%정도 성능이 좋다는 것을 보인다.

Prager[3], Clarke[4]는 고정길이 윈도우에 의한 단락검색 방법을 사용하였는데, [3]은 단락을 임의의 N개의 문장들로 구성하여 검색에 활용하였고 단락별 가중치는 tf*idf 값이 아닌 질문에 나온 키워드별 가중치를 이용해서 단락의 순위를 정했다. [4]에서는 질문의 개념을 이용해서 문서에서 부분 문자열을 추출하는데, 부분 문자열별 가중치는 질문의 개념과 일치되는 개념수, 문자열의 길이, 개념별 가중치값에 의해서 계산된다. 최종적으로 구해진 부분 문자열을 중심으로 고정된 크기의 단락이 추출된다. 가변길이 윈도우에 의한 단락검색 방법을 적용한 시도로는 이영신[5]등이 있다. 이 연구에서는 질의에 따라 정답을 포함될 가능성이 있는 후보영역의 크기가 다를 것이라는 가정 하에, 키워드를 포함하는 첫 문장과 끝 문장을 경계로 단락을 추출하여 검색하였다. 실험결과 문서검색이나 고정길이 단락검색보다 가변길이 단락검색의 재현율이 높음을 보였다.

의미적 실마리에 의해서 단락을 분류하는 연구는 Hearst[6], Kozima[7], Brants[8], Ponte[9] 등이 있다. 이러한 연구들의 기본 아이디어는 한 화제에서 다른 화제로 전환되는 부분을 어휘 사슬(lexical chains) 등을 활용하여 단락의 경계로 해석하는 것이다.

[6]은 문서를 세부토픽구조(subtopic structure)를 반영하여 응집성있는 다중단락 담화단위(multi-paragraph discourse units)로 분할하는 알고리즘인 TextTiling에 대해 기술한다. TextTiling은 어휘응집성의 척도로서 인접한 텍스트 블록의 쌍이 어휘적으로 얼마나 유사한가를 비교하고, 담화의 어느 지점에서 체인집합이 끝나고 새로운 체인집합이 시작되는지에 의해 세부토픽의 흐름을 결정한다. [7]은 어휘 응집성 프로파일(LCP : Lexical Cohesion Profile)이라는 구조를 이용하여 문서내 단락의 경계를 결정한다. LCP는 텍스트의 작은 윈도우(5에서 7개 단어) 안에 들어있는 단어들 사이의 유사도를 알아내기 위한 의미적 유사도 척도이며, 단어들간의 응집성을 표현하는 단어간 유사도는 의미망(Semantic Network)을 사용하여 계산하였다. [9]는 같은 주제의 문장들이지만 공통 키워드가 적을 경우, LCA(Local Context Analysis) 알고리즘을 통해 문장별 공통개념을 파악함으로써 같은 주제의 문장을 단락으로 구성한다.

이러한 대부분의 국내외 단락검색 연구는 문서의 문장 구조를 변형하지 않고, 특정 길이 또는 일정 범위의 문장들로 단락을 구성하기 때문에 한 단락에 다중 주제가 존재하거나 한 주제가 여러 단락에 걸쳐 기술된다. 따라서 질문의 주제를 활용한 단락검색 방법에는 적합하지 않다. 특히 문서 내에 공통 단어가 적고, 인접한 문장들이 개별적인 주제를 가지는 경우가 대부분인 백과사전에 적용하는 데는 한계가 있다.

3. 문장주제에 기반한 의미적 단락

백과사전은 사전에 수록된 어휘들에 대한 표제어 및 본문의 집합으로 구성되어 있다. 백과사전 본문을 구성하는 문장은 그 크기와 형식이 다양하다. 백과사전의 특성상 이미 종결된 사건이나 사실을 정리해서 요약하는 형식을 취하기 때문에 문장이 매우 간결하고 짧다. 반면 이론이나 현상을 설명하는 내용에서는 문장이 매우 복잡하고 다양한 구조를 취하게 된다.

백과사전의 본문이 갖는 또 다른 특성은 구성 문장들이 하나의 단락(paragraph)으로 나누어져 있으나 의미적인 응집도(semantic coherence)에 의하여 나누어진 것이 아니라는 점이다. 그림 1은 검색어 "박정희"에 대한 백과사전 표제어 및 본문 내용 중의 일부분이다. 예제문서가 두 단락으로 이루어져 있으나, 의미적인 응집도에 따라 구성된 것이 아니라 사용자가 읽기 편한 크기로 나뉘어져 있음을 알 수 있다. 특히 "인물" 도메인의 경우 그 인물에게 일어난 사건을 시간의 흐름에 따라 기술함을 알 수 있다.

박정희 朴正熙 [1917.11.14~1979.10.26]

경북 선산(善山) 출생. 가난한 농부인 박성빈(朴成彬)과 백남의(白南義) 사이에서 5남 2녀 중 막내로 태어났다. 1937년 대구사범학교를 졸업하고, 3년간 초등학교 교사로 근무하다가, 만주의 신경(新京 : 現 長春)군관학교를 거쳐 1944년 일본육군사관학교를 졸업하였으며, 8 ·15광복 이전까지 주로 관동군에 배속되어 중위로 복무하였다.

광복 이후 귀국하여 국군 창설에 참여하였으며, 1946년 조선경비사관학교(육군사관학교 전신) 제2기로 졸업하고 대위로 임관하였다. 그 후 육군포병학교장, 제5사단장, 제7사단장, 제1군 참모장, 제6관구 사령관, 육군본부 작전참모부장, 제2군 부사령관 등을 역임하였다. 1949년 사상 관련사건에 연루되어 군법회의에 회부된 적도 있었다. 당시 신문보도에 의하면 여수 ·순천사건 관련공산주의 협의자로 되어 있는데, 군법회의에서 무기징역을 언도받았으나 구명운동에 의해 복역은 면제되었다.
….

[그림 1] 백과사전 본문 예제

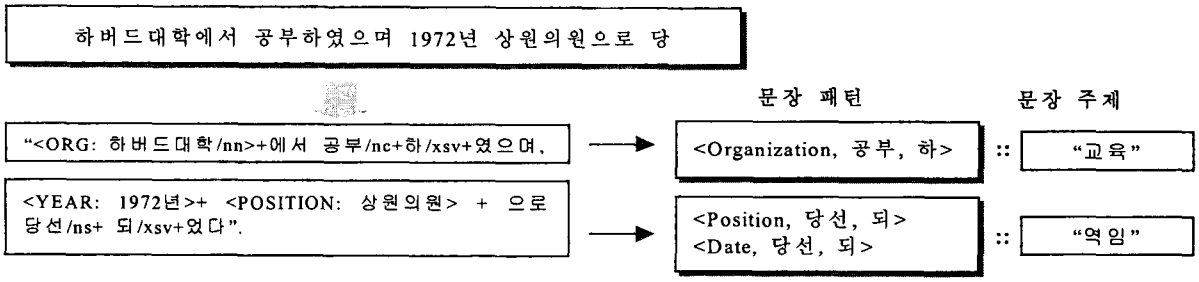
그림 1의 본문 첫 번째 문장을 보면, "*경북 선산(善山) 출생*"으로 매우 간결하며 짧다. 반면 세 번째 문장인 "*1937년 대구사범학교를 졸업하고, 3년간 초등학교 교사로 근무하다가, 만주의 신경(新京 : 現 長春)군관학교를 거쳐1944년 일본육군사관학교를 졸업하였으며, 8 15광복 이전까지 주로 관동군에 배속되어 중위로 복무하였다.*"의 경우, 여러 문장이 접속사에 의해 한 문장으로 구성되면서 공간적으로는 매우 인접한 위치에 있으나, 각 문장 "*1937년 대구사범학교를 졸업하고*", "*3년간 초등학교 교사로 근무하다가*", "*만주의 신경(新京 : 現 長春)군관학교를 거쳐 1944년 일본육군사관학교를 졸업하였으며*", "*8 15광복 이전까지 주로 관동군에 배속되어 중위로 복무하였다*"는 각각 "*졸업/교육/역임*"의 개별적인 주제를 가지고 있음을 알 수 있다.

본 논문에서는 이러한 특성을 반영하기 위하여 백과사전 문장에 주제를 할당하고, 이를 의미적인 단락으로 재구성하는 방법을 제안한다. 제안한 의미적인 단락 생성 방법은 문서를 주제별 단락으로 구성하여 색인/저장하는데 활용되며, 사용자가 질의한 주제와 부합되는 단락을 검색함으로써 질의/응답 시스템의 정확도를 향상시키는데 도움이 된다.

3.1 문장주제

의미적인 단락을 생성하기 위해서는 먼저 백과사전의 도메인의 특성을 반영하는 '문장 주제'를 선정해야 한다. 문장 주제란 어떤 도메인에 해당하는 문서 집합에서 공통으로 기술하는 내용이나 자주 언급하고 있는 사건 혹은 개념들의 집합을 의미한다. 앞에서 언급한 바와 같이 백과사전에 기술된 문장들은 하나 이상의 주제를 기술하고 있는 경우가 많다 '인물' 분야의 경우, 그림 1에서와 같이 시간의 추이에 따라 특정인에게 일어난 사건을 나열한 경우가 특히 많고, 사용자 질문 또한 특정인의 업적이나 중요한 사건에 관련된 경우가 대부분이다. 따라서 '인물' 분야에서 자주 출현하는 사건을 '문장주제'로 정의하고 이에 따라 의미적인 단락을 생성, 검색한다면 질의응답의 정확도 향상에 도움이 된다.

일반적으로 문장에 주제를 할당할 경우 문장에 등장하는 어휘에 기초하여 할당하게 된다. 또한 문서에 자주 등장하는 어휘일수록 그 문장의 주제를 할당하는데



[그림 2] 문장 패턴 자동 추출 결과

많은 영향을 미칠 수 있다. 본 연구에서는 백과사전에 자주 등장하는 어휘가 중요한 개념이라는 가정 하에, 어휘 빈도가 높은 단어를 대상으로 주제 카테고리를 설정함으로써 문장 주제를 정의하는데 드는 수작업을 최소화 하였다. 주제 카테고리로 설정될 후보명사를 선정하는 기준은 대상 도메인에서 추출한 모든 명사의 어휘 빈도를 계산한 후, 상위 30%의 어휘 빈도를 가지는 명사를 대상으로 하였다. 이렇게 추출된 명사를 대상으로 각 개념 간의 상 하위 관계를 분석하여 계층적 문장 주제를 선정하였다. 예를 들어 선정된 명사가 "이름", "본관", "별칭"인 경우, "이름"이 "본관"이나 "별칭"의 상위 개념이므로 "이름"을 대분류 주제로, "본관"이나 "별칭"을 소분류 주제로 선정하였다. 최종 선정결과, 8개의 대분류 주제와 35개의 개별 주제로 구성되었다(표 1 참고).

3.2 문장주제를 이용한 의미적 단락 생성

문장주제를 활용한 의미적 단락 생성이란 문장주제 할당기를 이용해 해당 주제 범주를 각 문장에 할당한 후, 이를 주제별로 문장을 재구성하는 과정이다. 학습기

를 통해 구축된 문장 주제 할당기의 주된 역할은 대상 문장에서 주제를 파악할 수 있는 단서를 인식하는 것인데, 여기서 가장 중요한 역할을 하는 것이 바로 '문장패턴'이다.

'문장패턴'이란 주어진 문장의 주제를 파악할 수 있는 가장 중요한 자질로, 핵심동사와 이를 중심으로 주변 문맥에 나타난 명사(NN) 와 개체명(NE) 태그 및 정답 유형(AT) 태그로 정의된다.

문장패턴 = 〈AT(NE, NN) , NN , V〉

예를 들어, 예문 "*영향/NN+을 받/V+았+다*"에서 추출되는 문장 패턴은 〈영향, null, 받〉이고, 예문 "*〈LOC : 하버드대학/NN〉+에서 공부/NN+하/xsv+였으며*"에서 추출되는 문장 패턴은 〈LOC, 공부, 하〉이다. 문장 패턴은 그림 2와 같이 형태소 분석 및 개체명 태깅, 정답 유형 태깅된 입력 문서로부터 자동으로 추출된다. 각 문장 패턴은 동사를 중심으로 주변 1개 명사 및 개체명(정답 유형)까지 추출된다. 그러나 접미사 "하다, 되다"의 경우, 바로 앞의 서술성 명사(이벤트성 명사)를 포함하여 추출한다. 이때, 문장 패턴 추출 경계(boundary)는 바로 앞 동사 이전까지로 한다. 즉, 일반 동사 '*받/V*'의 경우 〈풀리처상, null, 받〉으로 추출되며, 접미사 '하/xsv'의 경우 바로 앞의 서술성 명사까지 포함한 〈organization, 공부, 하〉를 문장 패턴으로 추출한다. 또한, 서술성 명사로 끝나는 문장, 예를 들어, "*〈LOC : 경상남도 영일〉 출생/NN.*"과 같은 문장의 경우, 서술성 명사를 동사패턴으로 적용한다.

이런 과정을 통하여 추출된 각 문장 패턴을 주제 카테고리로 분류하면 주제별 패턴군을 얻을 수 있다. 즉, 각 문장으로부터 추출된 〈organization, 공부, 하〉 패턴은 "교육" 주제로, 〈status, 당선, 되〉 패턴은 "역임" 주제로, 〈풀리처상, null, 받〉 패턴은 "수상" 주제로, 〈영일, null, 출생〉 패턴은 "출생" 주제로 할당된다.

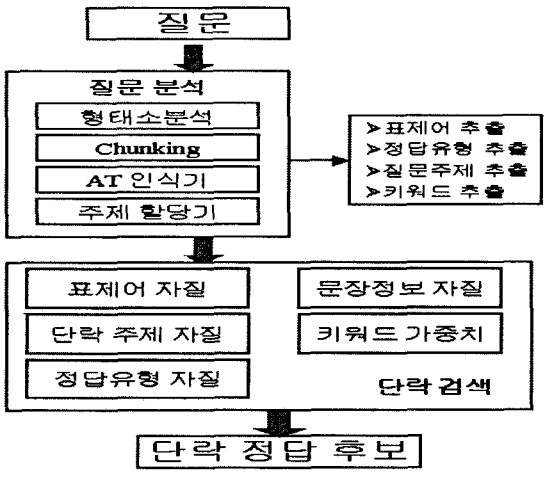
앞에서 언급한 바와 같이 문장패턴에서 주제를 부여하기 위한 가장 중요한 단서는 동사이다. 그러나 일반적으로 학습에 사용되는 문서의 양이 전체 문서집합에 비해 매우 적기 때문에, 학습문서에서 구축된 패턴에 출현한 동사의 자료부족문제가 발생한다. 본 논문에서는 학습문서에서 나타나지 않은 동사에 대한 적용 범위를 확장하기 위해 ETRI 동사 개념망을 활용하였다. ETRI 동사 개념망[1]은 각 동사가 하나의 개념노드로 상하위 관계로 연결되어 있으며, 유의어, 동의어, 사동/피동의 관계가 설정되어 있다. 예를 들면 '출생하다'의 경우 유의어로 '태어나다', '탄생하다' 등이 있고, '사망하다'의 유의어로는 '죽다', '급사하다' 등이 있다. 또한 '사망하다'의 피동형인 '암살당하다', '처형당하다' 등을 문장 패턴에 추가함으로써 학습문서에서 추출된 패턴의 적용범위를 넓힐 수 있게 된다. 실험결과, 학습문서에서 구축된 22,916개의 패턴을 38,214개로 확장시킬 수 있었다.

이렇게 추출된 문장패턴은 자동학습기를 통해 학습되어 문장주제 할당기에 반영된다. 문장주제 할당에 사용된 기계학습 알고리즘은 베이지언(Bayesian) 알고리즘을 사용하였는데, 문장에서 출현한 동사를 중심으로 해당 윈도우(패턴) 내에서 등장하는 어휘 확률값을 계산하여 최고 확률값을 갖는 주제를 문장주제로 할당하였다. 백과사전 문장은 한 문장에서 여러 사건 즉 주제들을 나열하고 있는 경우가 많으므로 각 문장을 단문단위로 분석하여 문장 주제를 할당한다.

4. 단락 검색

단락검색은 문장 주제에 의해서 재구성된 백과사전 본문의 의미적 단락을 대상으로 하며 질문과 가장 유사도가 높은 단락을 정답후보로 제시한다.

단락검색은 다음 그림과 같은 순서로 진행된다.



[그림 3] 단락검색 구성도

질문분석 모듈에서는 형태소분석, 청킹, AT인식, 단락주제할당 등의 언어분석결과를 이용해서 네 가지 정보를 추출하는데, 질문에 대한 표제어정보, 정답유형정보, 주제정보, 단일명사, 복합명사, 속격어구의 키워드정보들이다. 단락검색기에서는 이러한 네 가지 정보를 이용해서 단락검색을 수행한다.

1) ETRI 동사개념망은 현재 16,000개의 개별 동사로 구축되었으며, 2004년말까지 20,000개로 확장될 예정이다

4.1 자질

단락검색에 이용되는 자질은 다음과 같다.

- 표제어
- 단락 주제
- 정답 유형
- 문장 정보
- 키워드 가중치

검색에 미치는 영향력에 의해서 각 자질별로 가중치를 가지며, 자질별 가중치는 '표제어 〉 단락주제 〉 정답유형 〉 문장정보 〉 키워드 가중치' 순으로 부여된다. 질문과의 단락별 유사도는 위 다섯 가지 자질값의 합으로 구해진다.

4.1.1 표제어

표제어는 백과사전 도메인이 갖는 특징으로, 실험을 통해 특정 표제어에 대한 질문의 답은 해당 표제어의 본문에서 나올 확률이 높다는 가설이 검증되었다. 표제어 자질값은 질문에 표제어가 지정된 경우에만 부가되는 값으로, 질문 표제어와 동일한 표제어에서 나온 단락의 유사도에 표제어 자질값이 더해진다.

표제어 자질값이 계산되는 질문의 예는 다음과 같다.

- 질문 : **허블**이 천문학자가 되기 전의 직업은?
- 질문 표제어 : 허블

질문분석에서 잘못된 표제어를 줄 경우에는 단락검색의 성능을 저하시킨다. 따라서 질문분석에서 올바른 표제어를 추출하는 것이 중요하다.

표제어를 잘못 추출한 경우는 다음과 같다.

- 질문 : 다이쇼의 장자는 누구인가?
- 잘못된 표제어 : 장자
- 올바른 표제어 : 다이쇼

4.2.2. 단락주제

단락주제 자질은 질문의 주제와 동일한 주제를 가진 단락의 유사도에 부가된다.

단락주제 자질값이 계산되는 예는 다음과 같다.

- 질문 : 가우스의 출생장소는?
- 질문 주제 : 출생
- 단락 : **브룬스비크에서 노동자의 아들로 태어나** 빈궁한 가운데 성장하였지만, 어머니와 숙부의 노력으로 취학할 수 있었다.
- 단락 주제 : **출생**

4.2.3. 정답유형

본 시스템에서는 백과사전 인물 도메인에서 나타날 수 있는 정답유형을 70여개로 정의해서 사용한다. 정답유형은 계층구조로 이루어져 있는데 그림 4는 정답유형 계층구조의 일부분을 보여준다.

상위분류	하위분류	
Location	Country	
	Province	
	County	
	City	Capital City
	Geo Area	River
		Mountain
		Island
	Constellation	Planet
		Star
	Social Loc	

[그림 4] 정답 유형(AT)

정답유형 자질값은 사용자가 원하는 정답유형이 후보단락에 출현한 경우 부가된다. 정답유형의 계층구조를 활용해 정답유형 자질값을 구한 예는 다음과 같다.

- 질문 : 가우스가 태어난 곳은?
- 질문 정답유형 : Location
- 단락 : 브룬스비크에서 노동자의 아들로 태어나 빈궁한 가운데 성장하였지만, 어머니와 숙부의 노력으로 취학할 수 있었다.
- 단락 정답유형 : 브룬스비크(City), 노동자(Position), 아들(Position), 어머니(Position), 숙부(Position)

4.2.4. 문장정보

문장정보 자질값은 질문에 있는 키워드 중 두개 이상이 한 문장에 같이 나온 경우에 부여되는 값으로, 한 문장에 나온 질문 키워드별 가중치의 합에, 문장정보 가중치값을 곱해서 계산된다.

예를 들면 다음과 같다.

- 질문 : 미테랑이 한국을 방문한 해는?
- 질문 키워드(키워드 가중치) : 미테랑(2.4), 한국(1.2), 방문(0.2)
- 후보문장 : 미테랑은 1993년 9월 한국을 방문하여 기술교류를 논의하였다. …
- 문장정보 자질값 : (2.4+1.2+0.2)*(

여기서 (는 문장정보 자질값의 가중치로 실험을 통해서 구해진다.

4.2.5. 키워드 가중치

질문 및 단락에서 추출되는 키워드는 단일 명사, 복합 명사 및 속격 어구로, 속격 어구는 단일명사와 복합명사에 비해 높은 가중치를 가진다. 본 논문에서는 키워드별 가중치를 2-포아송 모델을 이용하여 구하였고 식은 다음과 같다.

$$W_d = \frac{tf_d}{K\left\{(1-b)+b\frac{\text{document length}}{\text{average document length}}\right\}+tf_d} \times \log\frac{N-df_t+0.5}{df_t+0.5}$$

5. 실 험

본 논문에서 제안하는 모델의 특징은 두 가지이다. 첫번째는 문장별 주제에 의해서 원문을 의미적 단락으로 재구성하는 것이고, 둘째는 질문의 주제와 단락의 주제를 고려해서 단락을 검색하는 것이다.

문장주제를 기반으로 생성된 의미적인 단락의 신뢰도를 평가하기 위해 문장주제 할당기의 성능을 분석하였고, 형식적인 단락이나 고정길이 단락을 대상으로 단락검색을 수행하는 이전의 연구 결과와 비교하기 위해 고정길이 단락검색 방법을 선택하였다. 단락검색의 대상은 백과사전의 전체 13개 도메인 중 '인물' 분야로 한정하였다.

5.1. 실험 환경

본 논문에서 사용한 평가셋은 ETRI 평가셋 1.0으로, 402개의 질의-정답 쌍, 정답 근거 문단 등으로 구성되어 있다. 이중 단답을 대상으로 하는 195개의 평가셋을 실험에 사용하였다.

의미적 단락 생성 신뢰도를 평가하기 위해서는 문서분류 시스템의 일반적 평가척도인 F-score를 사용하였고, 단락검색 성능을 평가하기 위해서는 역순위평균(MRR)과 사용자 정답만족도를 사용하였다. 역순위평균이란 사용자가 원하는 정답이 몇 번째에 나타났는가에 대한 순위를 평가하는 방법으로, 순위를 1/n의 가중치로 반영하는 척도로, TREC 등 여러 국제학회에서 사용하는 공통척도이다. 본 시험에서는 상위 5등까지의 순위를 평가하였다. 사용자 정답만족도란 5등 안에 정답이 포함될 경우에 시스템이 정답을 제시했다고 평가하는 것으로, 질의응답 시스템에 대한 사용자 정답만족도를 평가하기 위해 사용하였다.

5.2. 실험1 : 의미적 단락 생성 신뢰도

의미적 단락 생성의 신뢰도는 곧 문장 주제 할당기의 성능과 같다. 문장 주제 할당기의 성능을 평가하기 위해서 실험문장 1,131개의 문장을 사람이 수동으로 주제 분류하고, 개발된 시스템으로도 주제 분류하여 그 결과를 비교하였다. 어떤 문장에 대해 사람이 "활동, 참가,사망"의 주제로 분류하고, 시스템이 "죄, 조직, 이동, 사망"으로 분류하였다면, 재현율은 1/3=0.33 이고 정확율은 1/4=0.25 이다.

[표 1] 문장 주제별 할당 성능

대분류	중분류	소분류	F-score
출생			0.841
	국적		0.72
사망			0.944
이름			0.2
	본관		0.62
	별칭		0.3
활동			0.52
	업적		0.814
		참가	0.923
		설립	1
		조직	0.833
		기록	0.8
		연구	0.533
		주장	0.966
	연혁		0.75
		역임	0.808
		수상	0.778
	...		

실험 결과, 정확률 70.6%, 재현율 75.9%로 F-score 73.1%의 성능을 얻을 수 있었다. 표 1은 세부 성능 평가 결과를 나타낸 것으로, 문장 주제간 성능차이가 매우 큰 것을 알 수 있다. 이는 자료 부족(data sparseness)으로 인해 해당 주제를 대표하는 문장 패턴이 적게 추출됨에 따라 정규화되지 않았기 때문이다. 따라서 낮은 성능을 보이는 주제의 가중치를 낮춤으로써 단락 검색시 미치는 영향을 최소화하도록 한다.

또한 형태소분석기 및 정답 유형 태거의 성능이 높아지고 학습 문장 패턴에서 주제가 잘못 할당된 패턴을 교정하거나 중요하지 않은 패턴을 제거하는 등, 학습 데이터를 정제하면 충분히 신뢰할 만한 성능을 얻을 것으로 기대된다. 정제된 학습 문장을 대상으로 패턴을 추출하여 학습한 결과, 71.8%, 재현율 78.7%로 F-score 75.1%의 향상된 결과를 얻을 수 있었다.

5.3. 실험2 : 의미적 단락 검색 성능

5.3.1 고정길이 단락 검색과의 성능 비교

제안된 방법으로 생성된 의미적 단락의 검색 성능을 비교하기 위해 고정길이(3문장) 단락을 대상으로 키워드만 고려해서 검색한 결과를 비교하였다.

먼저 인물 도메인의 35개 주제분류에 의해 재구성된 의미적 단락을 대상으로 한 검색 성능과, 백과사전을 고정길이 단락단위로 단순히 키워드만 고려한 경우의 검색 성능을 비교 평가하였다. 평가 결과 표 2와 3과 같은 성능을 보였다. 실험결과로부터 주제를 고려한 의미적 단락 검색 방법이 고정 길이 단락을 검색하는 방법에 비해 사용자 정답만족도는 9%, MRR은 8% 정도 높은 성능을 보임을 알 수 있다.

5.3.2 문장 주제에 따른 검색 성능

주제가 의미적 단락 검색에 미치는 영향력을 알아보기 위해서 이번에는 의미적 단락을 주제를 이용한 경우와 이용하지 않은 경우로 나누어서 비교 실험하였다.

의미적 단락을 주제를 고려하지 않고 키워드만을 대상으로 검색한 결과는 표 4와 같다.

의미적 단락을 주제를 이용해서 검색한 결과인 표 2와 비교해 보면, 주제를 이용한 경우가 사용자 정답만족도는 10%, MRR은 6% 정도 우수함을 알 수 있다.

[표 2] 주제를 고려한 의미적 단락 검색

	1등	2등	3등	4등	5등
정답수	100	120	134	146	148
MRR	0.61				
사용자 정답만족도	0.76				

[표 3] 키워드만 고려한 고정길이 단락 검색

	1등	2등	3등	4등	5등
정답수	85	111	121	127	130
MRR	0.53				
사용자 정답만족도	0.67				

[표 4] 키워드만 고려한 의미적 단락 검색

	1등	2등	3등	4등	5등
정답수	96	108	122	127	128
MRR	0.55				
사용자 정답만족도	0.66				

5.3.3 의미적 단락 생성 성능 개선

앞 절의 결과와 같이 본 단락검색 모델에서 문장주제가 미치는 영향력은 매우 크다. [실험 1]에서 언급한 바와 같이 문장주제별 신뢰도 차가 크다. 낮은 신뢰도를 갖는 주제들은 단락검색기의 성능을 저하시키는 주요 요인이 된다. 따라서 주제별 정확도를 측정하고 임계치 이하의 주제들은 제거함으로써 단락검색에 미치는 오류를 최소화 하였다.

실험 결과 임계치 0.6 미만인 13개 주제가 제거되었으며, 남은 23개 주제범주를 이용한 의미적 단락검색의 성능은 표 5에 나와있다. 표 2과 비교해보면, 개선된 방법의 성능이 사용자 정답만족도는 0.5%, MRR은 5%가 향상되었는데, 이는 전체 정답 수는 1개 증가하였지만 모든 주제를 사용한 경우에 비해 높은 순위로 정답을 제시하였기 때문이다.

[표 5] 23 주제에 의한 의미적 단락 검색

	1등	2등	3등	4등	5등
정답수	116	128	138	148	149
MRR	0.66				
사용자 정답만족도	0.76				

6. 결 론

본 논문은 질문과 같은 주제를 가진 단락에서 정답을 추출하는 것이 질의응답 정확도를 향상시키는데 효과적이라는 가정 하에, 백과사전 본문을 문장별 주제에 의해서 의미적 단락으로 재구성하고 질문에서 적합한 주제를 파악하여 단락을 검색하는 방법을 제안하였다.

제안된 모델의 성능 분석을 위해 고정길이 단락검색과 비교하여 평가하였고, 실험결과 문장주제를 활용하여 의미적인 단락을 검색하는 방법이 사용자 정답만족도 및 MRR 모두 우수하다는 것을 알 수 있었다. 또한 문장 주제에 따른 검색 성능을 분석한 결과, 주제를 사용하는 경우가 그렇지 않은 경우에 비해 높은 성능을 보임을 알 수 있었다.

일반적으로 질의응답 시스템의 상용화를 위해서는 의미적 단락검색의 성능이 85% 이상의 정확도를 보여야 한다. 실험 결과 주제 할당기의 성능이 단락 검색의 성능에 가장 큰 영향을 주는 것으로 분석되었다. 앞으로의 연구계획은 다양한 방법에 의해서 주제를 정의 및 분류해 보고, 주제 할당기의 개선을 통한 단락검색기의 성능향상에 대해 실험해 볼 계획이다. 또한, 백과사전의 인물 분야뿐만 아니라 전체 도메인에 대한 적용 가능성을 모색해볼 예정이다.

참고 문헌

[1] Dan Moldovan, "Performance Issue and Error Analysis in an Open-Domain Question Answering System", in ACM Transactions on Information Systems (TOIS), Volume 21(2), 2003

[2] Salton, Allan, Buckley "Approaches to Passage Retrieval in Full Text Information Systems", in Proceedings of the 16th Annual International ACM- SIGIR Conference on Research and Development in Information Retrieval, pp49-58, 1993

[3] Prager, Brown, Coden, "Question-Answering by Predictive Annotation", in Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp184-191, 2000

[4] Clarke, Cormack, Lynam, Li, McLearn, "Web Reinforced Question Answering (MultiText Experiments for TREC 2001), in the Proceedings of Text Retrieval Conference (TREC), 2001

[5] Young-Shin Lee, Young-Sook Hwang, and Hae-Chang Rim, "Variable Length Passage Retrieval for Q&A System", in Proceedings of the 14th Hangul and Korean Information Processing, pp259-266, 2002

[6] Hearst, "Multi-paragraph Segmentation of Expository Text", in Proceedings of the 32nd Annual meeting of the Association of Computational Linguistics, 1994

[7] Hideki Kozima, "Text Segmentation Based On Similarity Between Words", in Proceedings of the 31st Annual meeting of the Association of Computational Linguistics, 1993

[8] Thorsten Brants, "Topic-Based Document Segmentat- ion with Probabilistic Latent Semantic Analysis", in Proceedings of the 11th International Conference on Information and Knowledge management, 2002

[9] Ponte, Croft, "Text Segmentation by Topic", in Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp113-125, 1997

[10] 황이규, 김현진, 장명길, "질의응답 기술개발", 정보처리학회지, 제11권 제2호, 2004