질의생성 모델을 이용한 전자우편 질의응답 시스템

장정선[0] 김상범 서희철 임해창
고려대학교 컴퓨터학과
{jszang, bewise, hcseo, rim}@nlp.korea.ac.kr

An E-Mail Question Answering System using Question Generation Model

Jeong-Sun Zhang[O] Sang-Bum Kim Hee-Chul Seo Hae-Chang Rim
Dept. of Computer Science and Engineering, Korea University

요 약

전자우편과 같이 일정한 질의 형식을 가지고 있는 긴 자연어 질의에 대해서 사용자 질의 단어에 가중치를 부과하는 방법과 질의에 대한 정답을 기존의 질의응답 집합에서 유사한 질의를 검색하여 그 정답을 사용자에게 제공하는 전자우편 질의응답 시스템을 제안한다.

사용자의 긴 자연어 질의가 주어지면 질의의 범주와 문장의 중요도 정보를 이용하여 질의에서 사용된 단어가 주제어로 쓰였을 확률을 계산하고, 계산된 확률에 기반하여 중요도를 할당하는 질의생성 모델을 제안한다. 또한 사용자 질의와 기존에 문의되어진 전자우편 질의의 유사도를 단어의 빈도를 고려한 어휘유사도, 한글 시소러스(Thesaurus)를 이용한 의미유사도와 본 논문에서 제안한 질의생성 모델을 이용한 주제 유사도를 이용하여 계산한다. 실험을 위하여 실세계에서 사용 중인 질의응답 집합을 이용하여 실험을 하였으며 각 유사도 계산 방법의 기여도를 비교 평가하고 제안한 질의생성모델이 성능향상에 미치는 영향을 평가하였다.

1. 서론

인터넷이 활성화되면서 통신채널 역시 인터넷 중심으로 변화하기 시작했고, 또한 통신수단으로서의 전자우편(E-Mail)의 사용도 급속도로 높아지고 있다. 전자우편이란 컴퓨터 통신망을 이용하여 컴퓨터 사용자간의 편지나 여러 정보를 주고받는 개인 통신 방법을 말한다. 전하고 싶은 내용의 편지나 컴퓨터에 수록된 자료를 다른 사람에게 보낼 수도 있고 받을 수도 있어 우편과 매우 유사한 특징을 지니며 전화와 같이 상대방에게 빠르게 편지나 자료를 전달할 수 있어 편지의 기록성과 전화의 즉시성이란 장점을 동시에 가지는 개인 통신 방법이다.

이러한 장점으로 인하여 전자우편은 특정분야에서 게시판을 대신하여 질의응답 방법으로 사용되고 있으며, 특히 금융이나 의료 등의 개인적인 정보에 대한 질의응답 방법으로 활발하게 사용되고 있다. 전자우편을 이용한 질의응답은 게시판을 이용한 질의응답에서와 마찬가지로 유사한 질의가 빈번하게 문의되어지고 있지만 게시판과 같이 정보 공유가 이루어질 수 없어 전자우편 질의응답에 많은 시간과 인력의 낭비가 발생하고 있다. 이러한 낭비를 방지하기 위하여 이미 문의되었던 전자우편 질의와 정답(이하 전자우편 집합)에 대한 정보를 검색하여 사용자의 새로운 전자우편 질의(이하 사용자 질의)에 대한 답을 제공하는 시스템을 전자우편 질의응답 시스템이라 한다.

전자우편 질의응답 시스템의 사용자 질의는 일반적인 정보검색 시스템이나 질의응답 시스템에서 자연어 질의나 문서로써 표현되는 사용자 질의와 다르게 전자우편으로 표현되기 때문에 전자우편의 특징을 지닌다. 전자우편 질의응답 시스템의 사용자 질의의 특징을 살펴보면 다음과 같다.

첫 번째, 전자우편 질의응답 시스템의 사용자 질의는 항상 두 문장 이상으로 구성된다는 문서적 특징을 지닌다. 두 번째, 전자우편 질의응답 시스템의 사용자 질의는 일반 편지의 형식을 지닌다. 편지의 경우 일반적으로 인사말, 본문, 맺음말이라는 형식을 가지고 있으며 전자우편의 경우 역시 이러한 특징을 지니며 모든 문장이 사용자가 묻고자 하는 질의를 표현하고 있지는 않다. 세 번째, 전자우편 질의응답 시스템의 사용자 질의는 묻고자 하는 주제가 존재한다. 전자우편으로 표현되는 사용자 질의는 사용자의 정보 요구를 표현하고 있다.

본 논문에서는 사용자의 질의를 사용자의 정보 요구에 의해서 생성되었다고 가정하고 사용자의 정보 요구를 직접적으로 표현할 수 있는 주제어를 추론하는 질의생성 모델을 제안한다. 질의생성 모델을 통하여 사용자

질의를 구성하고 있는 단어에 대해서 전자우편 질의를 생성할 확률을 계산하고 확률 분포에 따라 각 단어에 가중치를 할당한다. 이러한 가중치를 단어의 주제 가중치라 하며 주제 가중치를 사용하여 사용자 질의에서 사용된 모든 단어의 가중치를 부여하는 방법을 제안한다. 또한 사용자 질의와 전자우편 집합의 질의와의 의미유사도를 구하기 위해서 한국어 시소러스(Thesaurus)를 사용하여 단어와 단어간 거리 정보를 구하고 거리 정보를 이용하여 의미 유사도를 계산하는 방법을 제안한다.

2. 관련 연구

현재까지 사용자의 질의에 대한 적절한 정답을 내주는 질의응답에 대한 연구는 계속 진행되어 왔다. 정보검색은 사용자의 질의에 대해서 사용자가 요구하는 정보가 포함되어 있을 가능성이 높은 문서를 추출하는 것을 목표로 하고 있으며 단어 빈도나 문서 길이 등의 정보를 사용하여 사용자 질의와 문서의 유사도를 계산한다. 따라서 전자우편과 같이 사용자 질의가 상대적으로 긴 자연어 질의일 경우 사용자 질의 단어에 대한 가중치 부여 방법이 추가적으로 연구되어야 한다[6,9].

또한 1999년부터 TREC[1]에서는 사용자의 간단한 자연어 질의에 대해서 사용자가 요구하는 정보를 포함하는 제한된 크기의 단어나 구 등의 형태로 사용자에게 제공하는 것을 목표로 하는 질의응답 분야에 관한 연구가 현재까지 진행되고 있다. TREC에서 진행되는 질의응답 시스템은 방대한 문서집합에서 사용자의 요구에 대한 정확한 답을 찾는 시스템이다. 즉, 정보검색과 같이 질의에 대한 결과로 사용자가 요구하는 정보가 가장 많이 포함되어 있을 가능성이 있는 문서를 제시해 주는 것이 아니라 좀 더 사용자의 요구에 가까운 형태인 단어나 제한된 크기의 구의 형태로 제시해 주는 것이다. 질의응답 연구에서는 사용자의 질의가 단어열로 주어지는 것과는 달리 자연어 질의가 사용자 질의로 주어지기 때문에 TREC의 질의응답에 관한 연구는 이러한 사용자 질의를 분석하여 어떠한 정보를 추출할 것인지에 관한 질의분석 단계, 이렇게 추출된 정보를 가지고 방대한 문서집합에서 사용자가 요구하는 정보를 가지고 있을 가능성이 높은 문서, 단락, 문장 등을 어떻게 추출할 것인지에 관한 문서 검색 및 분석 단계, 그리고 이렇게 검색된 후보 문서, 단락, 문장에서 사용자가 요구하는 정확한 정답을 추출할 것인지에 관한 정답 추출 단계로 구분되어 연구가 진행되고 있다. 사실에 기반한 질의가 사용자 질의로 들어오며 광범위한 문서집합에서 사용자가 요구하는 정확한 정답만을 결과로 제출해야 하기 때문에 효과적으로 검색공간을 축소시킬 수 있는 방법으로 질의의 의도를 분석하여 질의 범주를 할당하는 방법 등에 관한 연구와 정보 검색 단계에서 검색 결과로 문서가 아닌 단락이나 문장을 제시하는 단락 검색 모델에 대한 연구, 그리고 정답 후보를 효과적으로 추출할 수 있는 방법 등에 대한 연구가 진행되고 있다. 이러한 TREC의 질의응답에 관한 연구는 정답 추출에 초점을 맞추고 있다. 그러나 전자우편 질의응답에 관한 연구는 사용자 질의와 유사한 전자우편 문서를 제시하는 것을 목표로 하고 있어 목표로 하는 결과가 서로 다르다. TREC의 질의응답에 관한 연구 역시 정보 검색 모델에서처럼 전자우편 질의응답 시스템과 같이 사용자 질의로 문서 단위의 질의 분석에 대한 연구가 요구된다[1,5].

또한 FAQ 자동응답 시스템은 자연어로 구성된 사용자 질의에 대한 결과로 사전에 구축된 FAQ 파일에서 질의와 유사한 FAQ를 검색하여 결과로 제시한다. 이러한 연구 분야는 우선 사용자의 질의로 자연어 질의가 들어온다는 점에서 TREC의 질의응답과 유사하다고 할 수 있으며, 결과로 유사한 문서를 추출한다는 점에서는 정보검색 연구방향과 유사하다고 할 수 있는 연구 분야이다.

전자우편 질의응답 시스템 역시 기존에 존재하고 있는 전자우편 질의와 정답의 쌍을 별도의 변형이나 추출, 가공 없이 사용자의 자연어 질의에 대한 정답으로 제공한다는 점에서 FAQ 자동응답 시스템과 유사하다.

FAQ 자동응답 시스템에서 사용자 질의와 기존 FAQ 파일의 유사도를 어떤 정보를 이용하여 계산하는지에 따라 어휘정보만을 사용하여 사용자 질의와 기존 전자우편 쌍과의 유사도를 계산하는 방법[8,9]과 어휘정보 외에 어휘의 의미정보까지 고려하여 유사도를 계산하는 방법[3]으로 나누어 질 수 있다.

어휘정보만으로 유사도를 계산하는 방법은 FAQ 자동응답 시스템은 같은 주제를 가진 영역에서 질의응답이 이루어지기 때문에 의미적인 어휘확장이나 의미 유사성을 계산하는 것은 불필요하다고 가정하였다. 주어진 단어를 가지고 비슷한 유사 부류를 찾아내어 추가적인 어휘 정보를 이용하였으며, FAQ 파일과 사용자 질의의 유사도는 단순히 사용자 질의와 FAQ 파일간의 일치 정도를 가지고 계산하였다[9].

또한 어휘의 의미정보를 대신하여 모든 어휘를 필수어, 선택어, 불용어로 구분하여 서로 다른 어휘 가중치를 할당하여 사용자 질의와 FAQ 파일간의 유사도를 계산하는 방법을 제안하였고 이를 통해 패턴이 비슷한 단어의 부류는 유사한 단어로 가정하였다. 단어의 분류를 통해서 단어의 중요도를 고려하였고 단어의 분류를 통해서 어휘의 의미정보를 부분적으로 고려했으나, 단어를 구분하는 이유와 분류 방법이 명확하게 제시되지 못했다[8].

의미정보를 고려한 방법은 주어진 질의의 어휘적인 정보만으로 효율적인 검색을 하기 어렵기 때문에 단어의 의미정보를 활용한다. 의미정보를 사용하지 않더라도 충

---

1) NIST와 DARPA의 지원을 받는 세계적인 정보검색 평가 대회이며 1992년 TIPSTER의 text program의 한 부분으로 시작되었다.
문서검색 방법론들의 대규모 평가를 위한 기반을 제공함으로써, 정보검색 분야의 연구를 지원하고 성능을 향상시키며 기술전이를 용이하게 하는 것을 목표로 하고 있다. 본 논문과 관련된 질의응답 분야는 1999년 TREC-8에서 처음으로 시도되었다.

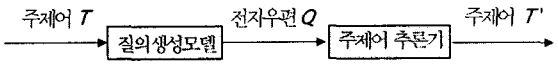
분히 큰 문서집합에서는 단어의 어휘 정보만으로 의미정보를 사용하는 효과가 있지만 문서의 크기가 상대적으로 작은 FAQ 자동응답의 경우 의미정보를 사용하지 않고서 단어의 의미정보를 획득할 수 없다고 보고 FAQ 자동응답 시스템의 경우 의미정보를 고려해야 한다고 제안하였다[3]. 의미정보를 고려하기 위하여 워드넷의 거리정보를 사용한 의미 유사도를 구하는 방법을 제안하였다. 의미 유사도 외에 단어의 빈도와 단어가 출현한 FAQ 파일의 빈도를 고려한 어휘 유사도와 FAQ 파일에서 질의의 출현 정도를 고려한 적용 유사도를 사용하였다. 이러한 방법은 워드넷의 거리 정보를 사용하여 의미 정도의 차이를 주어서 단어간의 분별력을 향상시켰고 FAQ 파일 빈도를 사용하여 단어의 중요도를 고려하였다. 하지만 의미 유사도를 구할 때 사용자 질의와 FAQ 파일에 나타난 모든 단어를 고려한 유사도는 사용자 질의나 FAQ 파일에 다의어가 존재할 때 전체적으로 의미 유사도가 상승하는 문제가 존재한다. 또한 사용자 질의가 여러 문장으로 구성되었을 경우 사용자 질의에 대한 가중치를 부여할 수 있는 방법이 없다.

3. 질의생성 모델

전자우편 질의응답 시스템의 목표는 사용자 질의와 유사한 전자우편 질의를 찾아내는 것이다. 전자우편 질의응답 시스템은 입력으로 사용자의 전자우편 질의가 주어지고 전자우편 집합에서 사용자 질의와 유사한 전자우편 질의를 검색하여 결과로써 제시하는 시스템이다.

전자우편 질의생성 모델은 전자우편 질의를 주제어에 의한 생성이라는 관점에서 해석하는 입력으로 주제어, 결과로 전자우편이 주어지는 언어모델이다.

예를 들어 질의 생성 모델을 설명하면 "대출 상품"에 대해 궁금한 것이 있어 전자우편 질의를 작성하는 사람이 있다고 가정을 해보자. 작성된 전자우편은 인사말도 있고, 자기소개도 있고, 기타 여러 문장으로 이루어져 있지만 생성 관점에서 바라본다면 작성자는 대출 상품에 대해서 궁금한 점이 있고 이를 설명하기 위해서 연관된 단어를 하나씩 추가한 것으로 볼 수 있다. 질의 생성 모델은 이러한 관점에서 전자우편 질의에서 주제어를 추론하는 모델이다. 이 과정은 [그림 1]과 같다.



[그림 1] 질의생성 모델

전자우편 Q는 주제어 T에 의해서 생성되고, 주제어 추론기를 통하여 주제어 T'이 추론된다. 추론된 주제어 T'는 다음 (수식 1)과 같이 추론될 수 있다.

$$T' = \arg\max_{w_t} p(w_t|Q) = \arg\max_{w_t} p(w_t)p(Q|w_t) \quad (1)$$

w_t는 주제어, $p(w_t)$는 단어의 주제확률, $p(Q|w_t)$는 주제어에 의한 전자우편 질의생성 확률을 의미한다.

3.1 범주정보를 이용한 단어의 주제확률

전자우편을 생성하는 주제어는 전자우편에 출현한 단어 중에서 단어의 주제확률과 질의생성확률의 곱을 최대로 하는 단어가 된다. 주제확률은 말 그대로 문서에 출현한 특정단어가 문서의 주제어일 확률이다.

아무런 정보가 주어지지 않은 상태에서는 (수식 2)와 같이 단어의 빈도를 이용하여 주제생성 확률을 계산한다.

$$p(w_t) = \frac{Count(w_t)}{\sum_t Count(w_t)} \quad (2)$$

하지만 전자우편 질의범주가 주어진다면 (수식 1)에서 사용된 주제생성 확률은 범주의 발생 확률과 범주와 주제어가 함께 발생할 확률을 이용하여 (수식 3)처럼 계산된다.

$$p(w_t) \approx p(w_t|c) = \frac{p(c, w_t)}{p(c)} \quad (3)$$

w_t는 주제어, c는 전자우편 질의범주, $p(c)$는 범주의 발생확률, $p(c, w_t)$는 범주, 주제어가 함께 발생할 확률, $p(w_t|c)$는 범주가 주어졌을 때 주제어가 발생될 확률

전자우편 집합에 있는 전자우편 질의에 질의범주가 할당되어 있다면 질의범주를 이용하여 범주의 자질단어를 추출하고 자질단어의 가망 비율(Likelihood Ratio)을 계산하여 가중치를 할당할 수 있다.

(수식 3)에서 범주와 주제어가 함께 발생할 확률 $p(c, w_t)$는 전자우편 집합에서 학습된 범주 자질 가중치를 이용하여 (수식 4)와 같이 계산된다.

$$p(c, w_t) = \frac{cw(c, w_t)}{\sum_t cw(c, w_t)} \quad (4)$$

여기에서 $cw(c, w_t)$는 범주의 자질단어의 가망비율,

$$cw(c, w) = \frac{p(c|w)}{p(\sim c|w)}$$

단어가 주어졌을 때 범주의 발생확률은 다음과 같이 빈도를 이용하여 계산한다.

$$p(c|w) = \frac{\delta + Count(c, w)}{\delta|V| + Count(w)}$$

$Count(c, w)$는 범주와 단어가 같이 출현한 빈도, $|V|$는 전체 어휘의 크기, δ는 최소 발생 빈도를 의미하며 $0 < \delta \leq 1$사이의 값을 갖는다.

또한 범주의 발생 확률 $p(c)$는 (수식 5)와 같다.

$$p(c) = \frac{Count(Q, c)}{Count(Q)} \quad (5)$$

$Count(Q, c)$는 특정 질의범주 c에 해당되는 전자우편 질의 수, $Count(Q)$는 전자우편 전체 집합의 수이다.

3.2 주제어에 의한 전자우편 질의생성 확률

전자우편은 두 문장 이상으로 구성되고 편지의 형식을 갖는다는 특징으로 인하여 전자우편을 구성하는 문장은 전자우편의 주제에 대해 표현하고 있지 않다.

예를 들면 전자우편 질의에서 "안녕하세요"란 문장과 "대출 자격 조건이 궁금합니다"라는 문장의 주제에 대한 표현력의 차이는 직관적으로도 알 수 있다.

전자우편 질의가 n개의 문장으로 구성되어 있다고 하면 전자우편 질의 Q는 다음과 같이 표현할 수 있다.

$$Q = s_1, s_2, s_3, \ldots, s_{n-1}, s_n = s_{1,n}$$

또한 전자우편을 구성하고 있는 각 문장의 발생이 서로 독립이라고 가정하면, (수식 1)의 주제어에 의한 전자우편 질의의 생성 확률 $p(Q|w_t)$는 (수식 6)과 같다.

$$p(Q|w_t) \approx p(s_1|w_t)p(s_2|w_t)\ldots p(s_n|w_t) = \prod_{i=1}^{n} p(s_i|w_t) \quad (6)$$

또한 전자우편의 i번째 문장 s_i가 주제어 w_t에 의해서 생성될 확률 $p(s_i|w_t)$는 주제어가 전자우편에서 출현한 단어라고 가정하였으므로 전자우편에서의 빈도를 이용하여 다음과 같이 계산한다.

$$p(s_i|w_t) = \frac{\delta + Count(s_i, w_t)}{\delta|Q| + Count(w_t)}$$

주제어에 의한 전자우편 질의의 생성확률을 (수식 6)에 의하여 구하게 되면, 전자우편의 형식적 특징을 반영하지 못하게 된다. 예에서 확인하였듯이 편지의 형식적 특징으로 인하여 전자우편을 구성하고 있는 문장의 주제에 대한 표현력에는 차이가 존재한다. 따라서 주제어에 의한 전자우편 질의의 생성확률도 중요문장을 생성했을 경우의 생성확률과 중요하지 않은 문장을 생성했을 경우의 생성확률이 서로 달라야 한다. 이러한 문장의 표현력을 (수식 6)에 반영해야 전자우편의 특징을 반영한 질의 생성 확률이 계산될 수 있다. (수식 6)에 문장의 표현력을 반영한 수식은 (수식 7)과 같다.

$$p(Q|w_t) \approx p(s_{1,n}|w_t) = \prod_{i=1}^{n} p(s_i|w_t)^{sw(s_i)} \quad (7)$$

여기에서 $sw(s_i)$는 문장의 표현력. 단, $0 < sw(s_i) \leq 1$

3.3 통계적 정보를 이용한 문장 표현력

문장의 표현력을 구하기 위해서는 주어진 문장이 전자우편에서 얼마만큼의 비중을 지니고 있는지 알아야한다. 이는 전자우편을 구성하고 있는 문장이 전자우편에서 얼마나 중요한 문장인지 알아내는 것과 같은 문제이고, 문서에서 중요문장을 추출하는 연구는 문서요약 분야에서 계속 연구되어져 왔다. 일반적으로 문서요약에서 중요문장을 추출하기 위한 연구는 크게 언어학적 접근 방법과 통계학적 접근 방법으로 나누어 볼 수 있다.

언어학적인 방법을 이용한 중요문장 추출에서는 어휘사슬, 담화트리 등을 이용하여 문서의 의미 구조를 파악한 다음 중요 문장을 추출한다. 비교적 높은 성능을 보이지만 워드넷, 시소러스 등의 추가적인 언어지식이 필요하다는 문제점이 있다.

통계학적 방법을 이용한 중요문장 추출에서는 단어의 빈도, 제목, 문장의 길이, 문장의 위치, 실마리 단어나 구 등의 통계값에 근거하여 중요문장을 추출한다. 각각의 자질에 가중치를 부여하는 방법에서 휴리스틱이 사용된다는 단점이 있지만 언어학적 접근방법에 비해 별도의 추가적인 언어 지식이 필요하지 않고, 또한 구하기 쉬운 정보를 사용하는 반면에 상대적으로 좋은 결과를 보인다는 장점이 있기 때문에 본 논문에서는 이 방법을 사용하여 문장의 가중치를 부여하였다.

통계학적 접근방법에서 문장의 가중치를 부여하기 위해서 사용되는 자질은 다음과 같다.

첫째, 단어의 빈도는 문서에 나타나는 단어의 분포에 기반하여 자주 나타나는 단어들은 문서의 가장 중요한 개념을 표현한다고 간주하여 이러한 단어를 가지고 있는 문장은 중요 문장이 된다.

둘째, 문서의 중요한 문장은 문서의 제목이나 절의 제목에 사용된 단어를 포함하는 경향이 있으므로 이러한 문장을 중요 문장으로 한다.

셋째, 길이가 너무 짧은 문장은 중요 문장에 포함되지 않는 경향이 있으므로 이러한 문장은 중요 문장에서 제외한다.

넷째, 문장의 위치에 따라 그 문장의 중요도가 다르다. 위치별 중요도는 문서의 종류에 따라 차이가 있으며 이 중요도는 기계학습 기법에 의해 학습할 수 있고 본 논문에서 사용되는 질의는 편지의 특징을 띄고 있으므로 상대적으로 중앙에 위치한 문장이 중요 문장이 된다.

다섯째, 문서에 나타나는 단어나 구 중에서 중요 문장이 될만한 중요한 부분임을 표시하는 표지의 역할을 하는 단어나 구를 이용하여 중요 문장을 선택한다. 예를 들면, 은행관련 전자우편의 경우 "문의", "궁금", "대출" 등이 이에 속한다. 또한 긍정적인 역할을 하는 단어뿐만 아니라 "안녕", "감사", "답변" 등과 같이 부정적인 역할을 하는 단어나 구를 이용하여 해당 문장을 제외시킬 수도 있다.

통계적 정보를 이용한 중요문장 추출에서는 위의 자질을 이용하여 중요 문장 추출을 위한 문장의 가중치를 다음과 같이 계산하였다.

$$SenScore(s) = \alpha C(s) + \beta K(s) + \gamma L(s) + \delta P(s) + \varepsilon T(s)$$

$C(s)$는 실마리 단어, $K(s)$는 단어의 빈도, $L(s)$은 문장의 길이, $P(s)$는 문장의 위치, $T(s)$는 제목을 이용한 문장의 가중치이다[4].

통계적 정보를 이용한 문장의 표현력은 위에서 구한 문장의 가중치를 이용하여 다음과 같이 계산한다.

$$sw(s_i) = \frac{SenScore(s_i)}{\max_k [SenScore(s_k)]}$$

3.4 질의생성 모델을 이용한 단어 가중치 할당

질의생성 모델은 주제어로부터 전자우편 질의가 생성되는 과정을 보여주는 모델이다. 또한 주제어 추론을 통하여 주어진 전자우편의 주제어를 (수식 1)을 통하여 추론할 수 있으며, 앞에서 살펴본 질의 범주 정보와 문장 표현력을 이용하여 (수식 1)을 표현할 수 있다.

먼저 주제확률과 질의생성 확률은 (수식 3)과 (수식 7)에 의해서 계산되고 이를 (수식 1)에 적용하면 다음과 같다.

$$T^{r} = \arg\max_{w_t} \frac{p(c, w_t)}{p(c)} \prod_i p(s_i|w_t)^{sw(s_i)}$$

위 식에서 상수 $p(c)$는 w_t와 관계없는 상수이기 때문에 제거하여도 결과에는 변동이 없다. 따라서 위 식은 다음과 같이 쓸 수 있다.

$$T^{r} = \arg\max_{w_t} p(c, w_t) \prod_i p(s_i|w_t)^{sw(s_i)}$$

또한 (수식 4)에서 분모 또한 주제어 w_t와는 상관없는 상수이므로 삭제해도 결과에는 변동이 없다. 따라서 위 식은 다음과 같이 변형될 수 있으며 단어의 순위화에 이용된 점수(Score)를 주제점수라 한다.

$$T^{r} = \arg\max_{w_t} cw(c, w_t) \prod_i p(s_i|w_t)^{sw(s_i)} \quad (8)$$

질의생성 모델을 이용하여 주제어를 추론하기 위한 (수식 1)은 최종적으로 (수식 8)과 같이 변형된다. 전자우편의 주제어를 추론하기 위해서는 전자우편 질의집합에서 출현한 모든 단어에 대해서 주제어로 쓰였을 확률을 계산하여 주제어 추론을 해야 되지만 생성 결과로써 전자우편이 주어지기 때문에 추론해야할 주제어는 주어진 전자우편 단어에 포함되어 있다고 가정한다.

따라서 본 논문에서는 질의생성 모델에서 입력으로 주어지는 주제어는 주어진 전자우편 질의에 출현한 단어라 가정하고 출현한 단어를 대상으로 하여 주제어를 추론한다. 이러한 가정과 (수식 8)을 이용하여 전자우편에서 출현한 모든 단어에 대해서 주제점수를 계산할 수 있으며 최고의 주제점수를 갖는 단어가 전자우편을 생성한 주제어가 된다. 또한 각 단어의 부여된 주제점수는 각 단어가 전자우편 질의에서 주제어로 쓰였을 정도이며 이를 단어의 주제 가중치로 할당하고 (수식 9)와 같다.

$$tw(w|Q) = cw(c, w) \prod_i p(s_i|w)^{sw(s_i)} \quad (9)$$

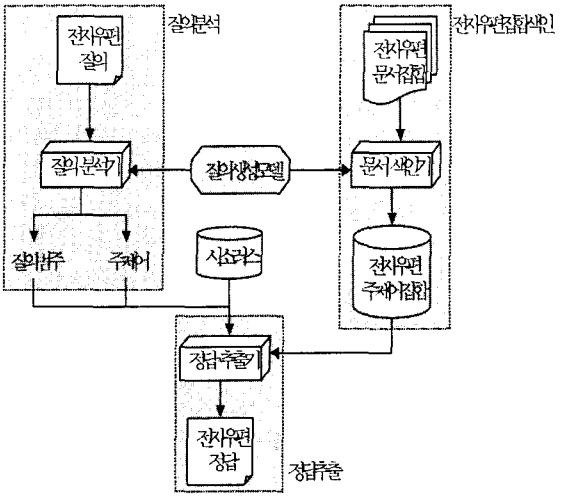
이렇게 계산된 주제 가중치는 사용자 질의와 전자우편 질의간 유사도 계산에 반영된다.

4. 질의생성 모델을 이용한 전자우편 질의응답 시스템

본 논문에서 제안한 전자우편 질의응답 시스템은 전자우편으로 된 사용자 질의를 입력으로 받아 기존의 전자우편 집합에서 사용자 질의와 유사한 전자우편 질의를 검색하여 결과로 제시하는 시스템이며, 전자우편 집합 색인, 사용자 질의 분석, 정답 추출의 3단계로 구분된다.

전자우편 집합 색인 단계에서는 전자우편 질의 집합을 분석하여 범주의 자질을 학습하고 전자우편 질의 단어의 주제 가중치 추출, 할당한다. 사용자 질의분석 단계에서는 사용자 질의범주의 할당과 사용자 질의 단어에 가중치를 부과한다. 정답추출 단계에서는 이러한 정보를 사용하여 분석된 사용자 질의와 전자우편 집합의 질의간의 유사도 계산을 통해서 사용자 질의에 가장 유사한 전자우편 질의의 정답을 사용자에게 제시한다.

본 논문에서 제안한 전자우편 질의응답 시스템의 전체 구성은 [그림 2]와 같다.



[그림 2] 전자우편 질의응답 시스템의 구성

4.1 전자우편 집합 색인 단계

전자우편 집합 색인 단계에서는 기존에 존재하는 전자우편의 질의와 정답에 관한 문서 중에서 전자우편 질의 집합의 범주를 이용하여 범주 자질을 학습하고 전자우편 질의에서 문서길이, 출현한 단어의 질의 생성 모델을 통한 주제 가중치와 단어 빈도 등의 정보를 추출하여 저장한다. 전자우편 색인을 위한 색인어 추출은 명사추출기를 사용하여 추출한다[2]. 불용어나 명사 추출 오류를 제거하기 위해서 문서 빈도가 임계값보다 낮은 단어는 색인 과정에서 제거한다.

범주 자질 학습은 범주별 전자우편 집합에서 명사추출기를 통하여 자질 후보 단어를 추출하고 가망 비율을 통해서 범주별 단어 가중치를 할당한다. 범주 자질 단어의 가망 비율은 다음과 같다.

$$cw(c, w) = \frac{p(c|w)}{p(\sim c|w)}$$

단어(w)가 주어졌을 때 범주(c)의 발생확률은 다음과 같이 빈도를 이용하여 계산한다.

$$p(c|w) = \frac{\delta + Count(c, w)}{\delta|V| + Count(w)}$$

$|V|$는 전체어휘의 크기, δ는 최소 발생 빈도($0 < \delta \leq 1$)

또한 범주의 발생 확률 $p(c)$는 (수식 10)과 같다.

$$p(c) = \frac{Count(Q, c)}{Count(Q)} \qquad (10)$$

Q는 전자우편 집합, $Count(Q)$는 전체 전자우편 수

4.2 질의분석 단계

사용자 질의분석 단계에서는 사용자 질의범주의 할당과 사용자 질의단어에 가중치를 부과한다. 사용자 질의 단어의 가중치는 질의생성 모델을 통하여 주제 가중치를 할당한다. 사용자 질의에 질의 범주를 할당하는 방법은 질의 집합 검색을 하여 검색 순위를 반영하여 범주 가중치가 가장 높은 범주를 할당하였고 이는 다음과 같다.

$$Class(q) = \arg\max_{c} \; ClassScore(c|D)$$

q는 사용자 질의, c는 질의범주, D는 순위화된 문서집합, $ClassScore(c|D)$는 순위화된 문서 집합이 주어졌을 때 범주 유사도를 의미한다.

범주가 할당된 순위화된 문서 집합이 주어졌을 때 범주의 유사도 $ClassScore(c|D)$는 다음과 같이 순위와 범주 발생 확률을 고려하여 계산한다.

$$ClassScore(c|D) = \sum_i \frac{f(d_i, c)}{Rank(d_i)} \frac{1}{p(c)}$$

$Rank(d_i)$는 문서 d_i의 순위, $p(c)$는 범주 발생 확률,

$$f(d_i, c) = \begin{cases} 1, & \text{if } Class(d_i) = c \\ 0, & \text{if } Class(d_i) \neq c \end{cases}$$ 이다.

4.3 정답추출 단계

질의 분석 단계에서 전자우편 질의 색인 단계와 질의 분석 단계에서 추출된 정보를 사용하여 사용자 질의와 전자우편 집합의 질의 사이의 유사도를 계산하여 가장 높은 유사도를 갖는 전자우편 질의에 해당되는 답을 정답으로 제시한다.

사용자 질의와 전자우편 집합 질의와 유사도를 측정하는 이유는 전자우편 집합의 질의와 정답 쌍과 사용자 질의의 유사도를 계산하는 것은 전자우편 정답에 부가적인 정보가 추가되어 성능을 저하시킬 가능성이 있기 때문이다. 이는 실험을 통하여 확인할 수 있다.

사용자 질의와 전자우편 집합 유사도는 어휘 유사도, 주제 유사도, 의미 유사도를 이용하여 계산한다.

단어와 문서간 어휘 유사도는 TREC-8의 Okapi 시스템에서 사용되었던 BM25방법에 의하여 가중치를 할당하였으며 다음과 같다[6].

$$LS(q_i, d) = \frac{tf_{q_i}}{k_1\left((1-b) + b\frac{dl_d}{avdl}\right) + tf_{q_i}} \log \frac{N - df_{q_i} + 0.5}{df_{q_i} + 0.5}$$

q_i는 i번째 사용자 질의 단어, d는 전자우편 집합의 질의 문서, N은 전체 전자우편 질의의 개수, tf_{q_i}는 사용자 질의 단어 q_i가 전자우편 집합 질의 d에서의 출현한 단어 빈도, df_{q_i}는 사용자 질의 단어 q_i가 출현한 전자우편 질의 문서의 빈도, k_1는 2, b는 0.75로 실험적으로 결정되었다.

사용자 질의 단어와 문서의 주제 유사도는 질의생성 모델을 통하여 계산된 주제 가중치(수식 9)를 사용하여 다음과 같이 계산한다.

$$TS(q_i, d) = \alpha \times tw(q_i|q) + (1-\alpha) \times tw(q_i|d)$$

q는 사용자 질의, α는 사용자 질의의 주제 가중치이다.

사용자 질의 단어와 문서의 의미 유사도는 한국어 시소러스를 사용하여 단어와 단어 사이의 의미 거리 유사도를 계산한 다음 단어와 문서의 의미 유사도를 다음과 같은 방법을 통하여 계산한다.

먼저 두 단어의 의미 유사도는 시소러스에서 두 단어의 평균 거리를 이용하여 다음과 같이 계산된다[3].

$$s(q_i, w) = H - (p \times \frac{H - L}{D})$$

w는 전자우편 집합 질의 문서 단어,
H, L는 두 단어 의미 유사도의 최대, 최소값,
p, D는 두 단어의 평균 의미 거리와 거리의 최대값

위의 수식과 단어의 주제 가중치를 사용하여 두 단어의 의미 유사도는 다음과 같이 계산된다.

$$ss(q_i, w_k) = tw(w_k|d)s(q_i, w_k)$$

최종적으로 사용자 질의 단어와 n개의 단어로 구성된 문서의 의미 유사도는 다음과 같이 단어간 의미 유사도의 최대값을 이용하여 계산된다.

$$SS(q_i, d) = \frac{tw(q_i|q)\ MAX_{k=1}^{n}\ (\ ss(q_i, w_k)\ )}{n}$$

정답 추출 단계에서 사용되는 최종적인 유사도는 어휘 유사도, 주제 유사도, 의미 유사도를 합하여 다음과 같이 계산한다.

$$SIM(q, d) = \frac{\alpha\sum_i LS(q_i, d) + \beta\sum_i TS(q_i, d) + \gamma\sum_i SS(q_i, d)}{\alpha + \beta + \gamma}$$

여기에서 α, β, γ는 각 유사도에 대한 가중치를 의미하고 $\alpha + \beta + \gamma = 1$이다.

전자우편 집합 질의와 사용자 질의간의 유사도는 위의 수식을 이용하여 계산되며 유사도 순으로 전자우편 집합 질의를 순위화하여 가장 높은 유사도를 갖는 전자우편 집합 정답을 사용자 질의에 대한 답으로 사용자에게 제공한다.

5. 실험 및 평가

5.1 실험환경 및 평가방법

본 논문에서 제안한 전자우편 질의응답 시스템은 금융관련 1422개의 전자우편 질의응답 집합과 100개의 사용자 전자우편 질의를 가지고 실험, 평가하였다. 전자우편의 질의범주는 전자, 대출, 카드, 예금, 기타, 외환, 신탁의 7가지 범주로 구성되어 있다.

실험집합으로 사용된 금융관련 1422개의 전자우편의 범주별 분포는 [표 1]과 같다.

범주	전자우편수	비율(%)	출현명사수	문서당명사수
전자	379	26.61	15,268	40.29
대출	371	26.05	17,412	46.93
카드	235	16.5	8,961	38.13
예금	206	14.47	8,177	39.69
기타	137	9.62	5,162	37.68
외환	70	4.92	3,341	47.73
신탁	24	1.69	1,034	43.08
총계	1422	100	59,355	41.74

[표 1] 범주별 실험집합 비교

전자우편 질의응답 시스템 평가를 위해 사용될 사용자 질의는 기존의 전자우편 질의를 유사한 질의로 재생성하여 구축하였다.

분명한 차이점 존재하는 서로 다른 전자우편 질의가 유사하다고 정의할 때, 두 전자우편의 차이는 동의어대체, 동의어추가, 동의어축소, 단어추가, 단어축소에 의해 발생되었다고 가정하여 평가 질의 또한 위의 다섯 가지 방법으로 기존의 전자우편 질의를 재생성하여 구축하였고, 재생성 하는 방법당 20개의 평가 질의를 구축하였다. 재생성 방법별 평가 집합은 [표 2]와 같이 구성되었다.

평가집합	개수	출현명사수	질의당 명사수
동의대체	20	617	30.85
동의추가	20	842	42.1
동의축소	20	325	16.25
단어추가	20	926	46.30
단어축소	20	342	17.1
전 체	100	3052	30.52

[표 2] 평가집합별 비교

전자우편 질의응답 시스템의 평가를 위해 평가 척도로 1위 재현율, 5위 재현율, R-Precision 등 3가지 척도로서로 비교, 평가하였다.

전자우편 질의응답 시스템의 특성상 1위의 중요성은 다른 유사 시스템과 비교할 수 없을 정도로 높다. 따라서 검색 결과 중에서 1위 문서만을 대상으로 시스템의 재현율(1-R)을 측정하여 비교하였다. 또한 5위까지의 결과 중에서 평가질의에 대한 재현율(5-R)을 구하여 이를 평가척도로 이용하였다.

전체적인 성능 비교를 위해서 순위화된 결과를 이용하여 정답이 출현한 순위에서의 재현율(R-P)을 사용하였고, 계산 방법은 다음과 같다.

$$R-Precision = Mean \sum_{i=1}^{N_q} \frac{1}{\text{정답순위}} \times 100$$

5.2 실험 및 평가

5.2.1 비교대상별 실험

본 논문에서는 사용자 질의와 기존의 구축되었던 전자우편 집합간의 비교 대상으로 전자우편의 질의를 사용하였다. 본 실험에서는 사용자 질의와 기존 전자우편간 유사도 측정 단위를 질의와 질의+정답간 유사도를 구하여 비교 평가하였다. 실험 결과는 [표 3]과 같다.

비교대상	질의수	1-R	5-R	R-P
질의+정답	100	51.0	70.0	59.6
질의	100	79.0	89.0	83.7

[표 3] 비교대상별 실험

[표 3]의 실험 결과에서 알 수 있듯이 사용자 질의와 기존의 전자우편 질의를 비교하여 유사도를 추정하는 방법이 사용자 질의와 기존의 전자우편 질의/답 쌍과 유사도를 추정하는 방법보다 더욱 효과적임을 알 수 있었다.

사용자 질의와 기존의 전자우편 질의와 유사도를 측정하는 게 더욱 효과적인 이유는 기존의 전자우편 질의에 대한 답에는 질의에 대한 정의 또는 설명으로 구성되어 있어 질의와는 상관없는 부가적인 단어의 사용이 많기 때문에 성능향상에 부정적인 영향을 끼쳤다.

5.2.2 질의 범주 정확도 실험

이번 실험은 (수식 23)을 이용하여 사용자 질의에 질의범주를 할당할 경우 범주 할당의 정확도에 관한 실험으로 결과는 [표 4]와 같다.

평가집합	질의수	맞게 할당된 질의수	정확도(%)
단어추가	20	17	85.0
단어축소	20	18	90.0
동의대체	20	19	95.0
동의추가	20	17	85.0
동의축소	20	17	85.0
전 체	100	88	88.0

[표 4] 질의 범주 정확도 실험

상대적인 평가는 아니지만 질의 범주를 결정하는 방법으로 매우 휴리스틱하고 직관적인 방법을 썼음에도 정확도가 90%가 가까운 높은 성능을 보였다.

5.2.3 주제 유사도 실험

본 논문에서 제안한 질의생성모델을 이용하여 계산한 주제 유사도가 성능 향상에 미치는 영향에 대해서 실험을 해보았다. 기본적으로 어휘와 의미 유사도를 측정하여 주제 유사도를 결합시킬 때 성능 변화를 비교하였다.

[표 5]는 주제 유사도가 미치는 영향을 분석한 결과이다.

평가집합	유사도 추정방법	질의수	1-R	5-R	R-P
단어추가	어휘	20	75.0	100.0	86.7
	어휘+주제	20	90.0	100.0	95.0
단어축소	어휘	20	70.0	90.0	79.4
	어휘+주제	20	75.0	90.0	81.9
동의대체	어휘+의미	20	70.0	75.0	72.4
	어휘+의미+주제	20	70.0	75.0	72.7
동의추가	어휘+의미	20	70.0	80.0	73.5
	어휘+의미+주제	20	75.0	85.0	79.9
동의축소	어휘+의미	20	75.0	90.0	83.1
	어휘+의미+주제	20	85.0	95.0	89.2
전 체	어휘	100	69.0	85.0	75.5
	어휘+의미	100	69.0	82.0	74.7
	어휘+주제+의미	100	79.0	89.0	83.7

[표 5] 주제유사도 실험

[표 5]에서 주제 유사도가 모든 평가집합에서 상당히 많은 기여를 함을 알 수 있다. 특히 단어 추가의 경우 주제 유사도가 성능 향상에 많은 도움을 줄 수 있음을 알 수 있다. 이는 단어의 수가 많으면 많을수록 단어의 분별력에 대한 가중치의 역할이 그만큼 커지기 때문이다. 따라서 사용자 질의가 긴 질의로 구성되는 전자우편 질의응답 시스템과 같은 경우에는 주제 유사도가 성능에 매우 큰 역할을 할 수 있을 것이다. 이는 단어 축소와 같은 평가 집합과 비교해보면 더욱 뚜렷하게 비교된다. 단어추가 평가집합 에서는 명확한 성능 향상을 관찰할 수 있었지만 단어축소와 같은 평가 집합에서는 상대적으로 낮은 성능향상을 보였다. 이는 단어가 축소된 경우 질의어 자체가 사용자의 정보 요구를 그대로 표현한다고 볼 수 있으므로 전부 질의어로써의 역할을 수행하고 있다는 것을 말한다. 그렇기 때문에 단어의 분별을 위해서 주제 가중치를 부여하는 방법이 그다지 성능향상에 도움을 주지 못함을 알 수 있다.

또한 동의어가 많이 쓰인 사용자 질의에 대해서는 단어추가 평가집합처럼 뚜렷한 성능향상을 보이지는 않지만 전반적으로 어휘유사도와 의미유사도만을 고려한 경우보다 성능이 향상되었음을 볼 수 있다. 특히 동의어가 사용되지 않은 평가집합과는 다르게 동의축소 평가집합에서도 성능이 향상됨을 볼 수 있는데 이는 의미 유사도 계산시에 주제유사도가 높은 단어에 높은 가중치를 주는 것이 유용함을 알 수 있다.

6. 결론 및 향후 연구

본 논문에서는 전자우편 질의응답 시스템을 위하여 사용자의 긴 질의에 대해서 질의생성 모델을 제안하였으며, 질의생성 모델을 통해서 질의 단어에 주제 가중치를 할당하고 이를 이용하여 사용자 질의와 전자우편 질의간 유사도를 추정하는 방법을 제안하였다. 본 논문에서 제안한 방법은 실험을 통해서 실제로 주제 유사도가 성능 향상에 기여함을 알 수 있었다.

본 논문에서는 주제 가중치를 할당할 때, 자료 부족 문제를 극복하기 위하여 가산적 평탄화 방법을 사용하였다. 좀 더 효율적으로 자료 부족 문제를 극복하기 위해서는 좀 더 다양한 평탄화 방법을 도입하여 전자우편 질의응답 시스템에 맞는 평탄화 방법을 찾는 연구가 계속 진행되어야 한다. 또한 사용자 질의의 범주를 결정하는 방법으로 기존의 전자우편의 검색결과만을 이용하는 방법을 제안하였으나 기존 연구와의 결합을 통한 범용적인 방법을 제안할 필요가 있으며, 다른 문서 분류 연구와의 비교 평가도 필요하다.

또한 전자우편 질의응답 시스템은 실세계에서 쓰이는 전자우편을 대상으로 하기 때문에 많은 오류를 포함하고 있다. 실세계에서 사용되는 전자우편은 철자 오류, 띄어쓰기 오류, 구어체 사용, 빈번한 미등록어의 사용으로 인하여 단어의 추출이 상당히 어려웠다. 따라서 실세계의 문서에 적용할 수 있는 단어 추출 모델에 대해서도 연구할 가치가 있을 것이다.

7. 참고문헌

[1]. 김수민, "시소러스 범주정보를 이용한 질의응답시스템", 제 12회 한글 및 한국어 정보처리 학술대회, pp179 ~ 183. 2000

[2]. 이도길, "명사 출현 환경을 고려한 빠른 색인어 추출 시스템",고려대학교 석사학위논문, 2000.

[3]. Burke, R., Hammond, K.; Kulyukin, V., Lytinen, S., Tomuro, N., and Schoenberg, S. "Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System", AI Magazine, vol. 18, no. 2: 57-66, 1997.

[4]. Edmundson, H. P., "New methods in automatic abstracting", Journal of the Association for Computing Machinery, 1969.

[5]. S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, P. Morarescu, "FALCON: Boosting Knowledge for Answer Engines", In the Nineth Text Retrieval Conference (TREC-9), 2000.

[6] Rebertson, S. E. et al., "Okapi at TREC-8", In the Eighth Text Retrieval Conference (TREC-8), 1999.

[7]. Eriks Sneiders, "Automated FAQ Answering: Continued Experience with Shallow Language Understanding", Papers from the 1999 AAAI Fall Symposium Technical Report FS-99-02, 1999.

[8]. Whitehead, S. D. "Auto-FAQ: an Experiment in Cyberspace Leveraging", Computer Networks and ISDN Systems, vol. 28, no. 1-2: 137-146. 1995.

[9]. Chengxiang Zhai, John Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval", SIGIR'01, September 9-12, 2001.