DOI: 10.3745/KIPSTB.2010.17B.4.327

질문대답 아카이브에서 어휘 연관성을 이용한 질문 분류

김 설 영[†] · 이 경 순[††]

요 약

보통 두 세 개의 어휘로 구성된 질문 분류에서 어휘의 다양한 표현으로 인한 어휘 불일치문제는 성능 저하의 주요 원인이다. 따라서 질문 분류에서 어휘 사이의 연관성을 반영하는 것이 필수적이다. 본 논문에서는 같은 범주의 질문-질문 쌍들에 대해 계산한 어휘 번역확률을 번역기반 언어모델에 반영하여 질문을 분류하는 방법을 제안한다. 실험에서 야후!앤써 질문대답 아카이브를 이용해서 전체 질문-대답 쌍들에 대해서 번역확률을 계산하는 것보다 같은 범주에 속하는 질문-질문 쌍들에 대해서 번역확률을 계산하는 것이 질문 분류에서 더 좋은 번역확률인 것을 증명한다.

키워드 : 질문 분류, 어휘 연관성, 번역기반 언어모델, 언어모델, 자질 추출, 질문대답 아카이브

Question Classification Based on Word Association for Question and Answer Archives

Xueying Jin[†] · Kyung-Soon Lee[††]

ABSTRACT

Word mismatch is the most significant problem that causes low performance in question classification, whose questions consist of only two or three words that expressed in many different ways. So, it is necessary to apply word association in question classification. In this paper, we propose question classification method using translation-based language model, which use word translation probabilities for question-question pair that is learned in the same category. In the experiment, we prove that translation probabilities of question-question pairs in the same category is more effective than question-answer pairs in total collection.

Keywords : Question Classification, Word Association, Translation-Based Language Model, Language Model, Feature Selection, Question and Answer Archives

1. 서 론

질문 분류는 사용자가 웹에 제출한 질문을 미리 정해져 있는 범주로 분류하는 것으로, 웹 검색에서 질문에 대한 분류정보를 이용해서 영역별 검색(Vertical Search)에 활용할 수 있다. 영역별 검색은 웹의 다양한 분야에 대한 검색결과를 보는 통합검색과는 달리, 질문에 해당하는 특정한 영역 내에서 검색함으로써 검색의 정확도를 높일 수 있다. 질문은 보통 'free books'처럼 두세 개 어휘로 이루어졌기 때문에 어휘 불일치문제(word mismatch problem)가 발생하여 일반적인 문서 분류보다 어렵다. 질문 분류에 관한 연구로 KDDCUP 2005[1]에서는 두 세 개의 어휘로 구성된 짧은 질문을 하나 이상의 범주로 분류하는 문제를 다루고 있다.

본 논문에서는 질문 분류의 어휘 불일치 문제를 해결하기 위해 번역모델을 이용하여 질문을 분류한다. 같은 범주내의 질문들은 그 범주를 대표하는 주제를 표현할 것이기에 야후!앤써(Yahoo!Answers)[2] 컬렉션의 학습집합에 대해서 같은 범주에 속하는 질문-질문(Q-Q) 쌍을 이용하여 번역확률을 계산하는 방법을 제안한다.

제안한 방법의 유효성을 검증하기 위해 언어모델을 이용한 분류기와 비교실험을 하였다. 질문 분류에서 번역확률 계산 방법에 따른 성능 변화를 관찰하기 위하여 어휘들 사이의 번역확률을 질문-대답(Q-A) 쌍을 이용해서 계산하는 방법과 각 범주에서 질문-질문(Q-Q) 쌍을 이용하여 번역확률을 계산하는 방법을 비교하였다. 질문 분류에서 범주 정보를 이용하여 범주 내에서 질문-질문 쌍으로 계산한 번역

※ 이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-531-2008-1-D00038). 또한 이 연구에 참여한 연구자는 2단계 BK21사업의 지원비를 받았음.
† 준 회 원 : 전북대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수
논문접수 : 2010년 2월 25일
수 정 일 : 1차 2010년 4월 19일
심사완료 : 2010년 4월 20일

확률이 분류 성능이 더 우수하다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구에 대해 기술하고, 3장에서는 어휘 연관성을 이용한 질문 분류에 대해 소개하고, 4장에서는 어휘 번역확률 계산방법에 대해 설명하고, 5장에서는 실험 및 평가, 6장에서는 결론에 대해 언급한다.

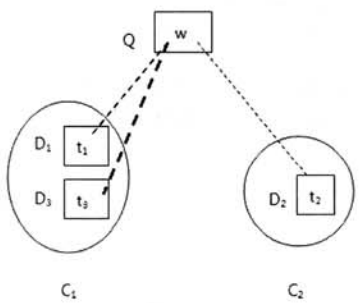
2. 관련 연구

정보검색에서 Berger과 Lafferty[3]는 어휘들 사이의 번역확률을 이용하여 IBM 모델 1로 문서들을 검색하는 방법을 소개하였다. Jiwoon[4]는 IBM 모델과 언어모델을 결합한 번역기반 언어모델을 제안하였다. 질문대답 아카이브에서 어휘 번역확률을 계산하여 질문대답 검색을 하여 비슷한 질문을 찾을 때 번역기반 언어모델이 언어모델에 비해 성능이 향상됨을 보였다. Xiaobing[5]은 [4]에서 제안한 번역기반 언어모델에서 번역확률을 계산하는 방법에 따른 성능 변화를 비교하였다. 또한 질문대답 검색에서 질문 문장뿐만 아니라 대답 문장에 대한 언어모델 검색결과를 반영하였다. 번역확률 계산방법은 질문과 대답을 쌍으로 묶어주는데 있어서, 질문-대답 쌍 또는 대답-질문 쌍 등 질문과 대답을 각각 번역부분의 원시로 하였을 때의 확률을 계산하는 방법과 질문-대답 쌍과 대답-질문 쌍을 하나의 풀(pool)에 넣고 전체에 대해 번역확률을 계산하는 방법 등에 대해 윈더 컬렉션(Wondir Collection)에 대해 실험하였다. 본 논문에서는 질문 분류를 위하여 같은 범주의 질문-질문 쌍들에 대해 번역확률을 계산하였다.

질문 또는 질의 분류는 이미 정해진 범주들에 대해서 새로운 질문 혹은 질의를 하나 이상의 범주로 분류하는 것이다. Cao[6]는 사용자가 검색엔진에 연속 올려놓은 인접한 질의들은 의미상 연관된 질의들이란 가정을 이용하여 질의A의 범주 정보를 이용하여 사용자의 의향을 파악하여 질의B의 범주를 판단한다. 이와 같이 문맥 정보를 이용하는 방법은 온라인 모드의 실제 분류 시스템에 큰 도움이 된다. Dou Shen[7]는 KDDCUP 2005 데이터 집합의 질의와 범주를 확장(enrich)하여 분류하였다. 질의와 목표 분류 체계 사이에 중간 분류 체계 Open Directory Project(ODP)[8]를 삽입하여 질의를 먼저 중간 범주에 분류한 후 중간 분류 체계와 목표 분류 체계가 갖고 있는 직접 매칭 혹은 확장 매칭에 의해 질의를 목표 범주로 분류하였다. 이 방법은 온라인 모드에서 목표 분류 체계의 범주가 변해도 질의를 중간 분류 체계에 분류할 수 있어 최종적으로 질의를 목표 범주로 분류한다. Yu[9]는 KDDCUP 2005 데이터 집합의 질의를 확장하여 분류하였다. 야후 분류 디렉토리(yahoo classification directory)의 어휘를 씨앗 어휘(seed word)로 구글에서 100개 요약문을 검색한다. TF-IDF를 이용한 벡터모델로 씨앗 어휘와 요약문의 유사도를 계산한 후 유사도가 가장 높은 요약문들을 자질추출을 위한 원시로 한다. 각 어휘의 벡터는 요약문의 기타 어휘의 가중치(weight)로 표현하며 코사인 유사도로 어휘들 사이의 유사도를 계산하여 관련어휘(relevant words)들을 순위화한다. 관련어휘들의 유사도에 임계값(threshold)를 주어 관련어휘로 질의를 확장하였다. Dell[10]은 SVM분류기가 질문 문법구조의 장점을 이용할 수 있는 트리 커널(Tree Kernel) 수식을 제출하여, 핵심 어휘에 가중치를 크게 주어 질문을 분류하였다.

3. 어휘 연관성을 이용한 질문 분류

본 논문에서 질문 분류는 검색을 통한 상위의 검색결과를 이용해서 검색 문서가 많이 속하는 범주로 분류를 하였다. 어휘 연관성을 반영한 분류 형태는 (그림 1)에 나타나있다. 범주 C_1에 문서 D_1, D_3이 있고, C_2에 문서 D_2가 있으며 각 문서에 포함된 어휘는 그림과 같다. 범주 C_1과 C_2의 문서들에 어휘 w가 존재하지 않기에 어휘매칭으로 문서들을 검색할 수 없다. 어휘 연관성을 이용하여 질문에 어휘 w가 존재하지 않지만 w와 연관 관계가 있는 어휘 t_i(i=1,2,3)가 있을 때 모델은 어휘 w와 t_i의 연관성을 이용해서 t_i를 포함한 문서들을 검색해 준다. 점선의 굵기가 어휘들 사이의 연관성을 확률로 나타낼 때 w는 t_3과의 확률이 제일 크고 그 다음으로 t_1과 t_2이다. 즉 $P(w|t_3)>P(w|t_1)>P(w|t_2)$이므로 문서의 검색 순위는 D_3, D_1, D_2 순위가 된다. 기존의 최근접 이웃 분류방법(k-Nearest Neighbors)으로 분류할 때 학습문서들 중에서 질문과 가장 유사도가 가장 높은 순위인 k개의 문서를 구하고 그들이 가장 많은 빈도를 보인 범주로 분류한다. k=3일 때 C_1에 속하는 문서가 두 개 있고 C_2에 속하는 문서가 한 개 있으므로 질문 Q는 C_1로 분류하게 된다.



(그림 1) 어휘 연관 정보를 반영한 분류 예

어휘 연관성을 반영하는 번역기반 언어모델(Translation-based Language Model)은 언어모델(Language Model)과 어휘 번역확률을 반영하는 IBM 모델 1[3]을 개선한 방법으로, 질문대답 아카이브에 대한 검색에서 언어모델에 비해 우수한 성능을 보였다[4]. 번역기반 언어모델을 질문 분류에 적용시키기 위해서는 어휘에 대한 번역확률을 획득하는 방법이 필요하다.

번역기반 언어모델의 기본이 되는 언어모델의 수식은 다음과 같다.

$$P(Q|D)=\prod_{w\in Q}P(w|D) \quad (1)$$

$$P(w|D) = (1-\lambda)P_{ml}(w|D) + \lambda P_{ml}(w|C) \quad (2)$$

$$P_{ml}(w|D) = \frac{freq(w,D)}{|D|},\ P_{ml}(w|C) = \frac{freq(w,C)}{|C|} \quad (3)$$

여기서 Q는 질문을 나타내고, D는 검색할 문서를 나타낸다. |D|와 |C|는 각각 문서 D에 나타난 어휘 개수와 켈렉션 C에 포함된 어휘 개수를 나타낸다. freq(w,D)와 freq(w,C)는 각각 어휘 w가 문서 D에 나타난 빈도수와 어휘 w가 컬렉션 C에 나타난 빈도수를 나타낸다. $P_{ml}(w|D)$는 어휘 w가 문서 D에서의 확률이다. 수식 (2)에서 파라미터 λ는 문서에서의 어휘의 확률과 컬렉션에서의 어휘의 확률에 대한 비중을 반영하는 것으로 학습을 통해 결정한다.

번역기반 언어모델은 어휘 번역확률 정보를 언어모델에 반영시키기 위해 수식 (2)를 아래 수식 (4), (5)와 같이 변형하였다.

$$P(w|D) = (1-\lambda)P_{mx}(w|D) + \lambda P_{ml}(w|C) \quad (4)$$

$$P_{mx}(w|D) = (1-\beta)P_{ml}(w|D) + \beta\sum_{t\in D} P(w|t)P_{ml}(t|D) \quad (5)$$

수식 (5)에서 P(w|t) 는 어휘 t가 w로 번역된 확률이고 $P_{mx}(w|D)$는 w와 w의 번역어휘 t를 함께 고려한 확률이다. 질문에 나타난 어휘 w가 문서에 나타나지 않았더라도 문서에 나타난 어휘 t와 w의 연관성을 이용하여 검색에 반영한다. 파라미터 β로 어휘 번역확률 부분의 중요도를 조절할 수 있다. 만약 β에 작은 값을 부여한다면 언어모델과 비슷하게 된다.

4. 어휘 번역확률 학습

일반적으로 어휘 번역확률은 기계번역(Machine Translation)에서 서로 다른 언어 쌍에 대해서, 예를 들어 한국어 문장과 이에 대한 영어 번역 문장을 이용하여 한국어 문장에 나타난 어휘가 영어문장에 나타난 어휘로 번역될 확률을 계산하는 것이다. 한국어 문장을 원시(source)로 보고, 이에 대한 영어 번역문장을 목적(target)으로 보고, 한국어 어휘에 대한 영어 어휘의 번역확률을 계산한다.

질문과 대답이 같은 언어로 구성되어 있는 질문대답 아카이브에서 어휘-어휘 사이의 번역확률을 얻기 위해서는 원시와 목적은 질문과 대답 중 임의로 어느 하나를 원시로, 다른 하나를 번역문(목적)으로 정할 수 있다. 어휘 번역확률은 두 어휘 사이 연관성의 중요도를 확률 값으로 나타낸다. 질문과 대답의 어휘 연관성을 아래 예와 같이 표현한다.

질문: Best way to loose weight?
대답: eat less exercise more

어휘 'loose'와 'eat', 'less', 'exercise' 사이, 'weight'와 'eat', 'less', 'exercise' 사이에 어휘적 연관성이 있다는 것을 알 수 있다.

본 논문에서 질문대답 아카이브에서 어휘 사이의 번역확률을 계산하는 방법으로 질문-대답 쌍을 이용하는 방법과 질문 분류를 위한 학습범주에서 각 범주에 속하는 질문-질문 쌍을 이용한 방법으로 어휘 번역확률을 계산하였다. 어휘 번역확률은 EM알고리즘을 이용한 GIZA++[11]를 이용하여 계산하였다.

4.1 질문-대답 쌍을 이용한 학습

질문 $q_1,q_2,\cdots,q_M$으로 구성된 질문 집합을 $Q=\{q_1,q_2,\cdots,q_M\}$라 하고, 대답 $a_1,a_2,\cdots,a_N$으로 구성된 대답 집합을 $A=\{a_1,a_2,\cdots,a_N\}$라 할 때, 질문과 대답으로 구성된 질문-대답 쌍 $(q,a)_i$는 질문을 원시로, 대답을 목적으로 본 것이고, 대답-질문 쌍 $(a,q)_i$는 대답을 원시로, 질문을 목적으로 본 것이다.

$P(w_i|w_j)$는 원시 어휘가 w_j일 때 목적 어휘가 w_i일 번역확률을 나타낸다.

$P(w_i,A|w_j,Q)$는 질문(Q)를 원시, 대답(A)를 목적으로 하였을 때의 번역확률이며 P(A|Q)로 표현하여 질문-대답 쌍 컬렉션 $\{(q,a)_1,...,(q,a)_n\}$으로 학습한다. $P(w_i,Q|w_j,A)$는 대답을 원시, 질문을 목적으로 하였을 때의 번역확률이며 P(Q|A)로 표현하여 대답-질문 쌍 컬렉션 $\{(a,q)_1,...,(a,q)_n\}$으로 학습한다.

P(A|Q)와 P(Q|A)를 결합하는데 두 가지 방법이 있다. 첫 번째 방법은 질문-대답 쌍과 대답-질문 쌍을 컬렉션 $\{(q,a)_1,...,(q,a)_n,(a,q)_1,...,(a,q)_n\}$에 모두 포함시켜 번역확률 $P(QA)_{pool}(w_i|w_j)$를 학습하는 것이다. 두 번째 방법은 P(A|Q)와 P(Q|A)를 각각 구한 후 선형 결합하는 방법인데 수식은 아래와 같다.

$$P(QA)_{lin}(w_i|w_j) = (1-\delta)P(w_i,Q|w_j,A) + \delta P(w_i,A|w_j,Q)$$

<표 1>에서 어휘 'lose'에 대해 질문과 대답 번역 쌍의

〈표 1〉 어휘 'lose'에 대한 질문-대답(Q-A) 쌍과 질문-질문(Q-Q) 쌍에서의 번역확률

질문-대답 쌍								질문-질문 쌍			
P(A|Q)		P(Q|A)		$P(QA)_{lin}$		$P(QA)_{pool}$		$P(Q\|Q)_{100}$		$P(Q\|Q)_{300}$	
lose	0.2392	lose	0.1967	lose	0.0984	lose	0.1211	lose	0.1726	weight	0.1048
weight	0.0902	eat	0.0772	eat	0.0386	eat	0.0480	weight	0.1564	lose	0.0788
day	0.0307	lost	0.0580	lost	0.0290	exercise	0.0376	fat	0.0182	diet	0.0180
loose	0.0242	exercise	0.0416	exercise	0.0208	weight	0.0375	diet	0.018	loss	0.0151
lost	0.0222	weight	0.0410	weight	0.0205	lost	0.0351	way	0.0138	fat	0.0149

구성에 따른 번역확률 계산방법으로 계산한 예를 나타냈다. $P(QA)_{lin}$의 δ는 0.5로 설정한 상태이다. $P(Q|A)$와 $P(QA)_{lin}$은 목적을 순위화한 후의 순위가 같다는 것을 볼 수 있다. 이런 결과가 나타나는 원인은 $P(Q|A)$의 번역 확률이 $P(A|Q)$보다 더 높기 때문이다. 예를 들어, 목적이 모두 'lose'일 때, $P(Q|A)=0.2388$, $P(A|Q)=0.2101$이다. $P(QA)_{lin}$을 제외한 나머지 세 방법으로 얻은 번역 어휘의 순위는 서로 다른데 $P(QA)_{pool}$방법으로 $P(Q|A)$와 $P(A|Q)$에서의 중요한 어휘를 모두 얻을 수 있다.

4.2 각 범주의 질문-질문 쌍을 이용한 학습

질문 분류에서 같은 범주에 속하는 질문들은 범주를 대표하는 어휘들이 포함되어 있으므로 어휘 사이의 연관성은 분류를 위한 중요한 정보가 될 수 있다. 따라서 각 범주에 속하는 질문-질문을 번역 쌍으로 구성하여 번역확률을 계산하였다.

$P(w_i,Q_p|w_j,Q_q)$는 Q_q를 원시로, Q_p를 목적($p\neq q$)으로 로 하였을 때의 번역확률이다. 어떤 한 범주에 속하는 질문들에 대해서 하나의 질문 Q_q에 임의의 개수의 질문 Q_p를 번역 쌍으로 만들 수 있다. 모든 범주의 번역 쌍들을 하나의 풀에 넣고 번역확률을 구한다. <표 1>에서 각 범주의 번역 쌍 개수를 100, 300 개로 하였을 때 어휘 'lose'의 번역확률의 일부를 나타냈다. $P(Q|Q)_n$은 n개의 Q-Q쌍을 만들었을 때의 확률을 표시한다.

5. 실험 및 분석

제안한 방법의 유효성을 평가하기 위해 야후!앤써 실험집합[2]을 이용하여 실험하였다.

5.1 야후!앤써 실험집합

야후!앤써 집합은 Q&A 아카이브로서 대량의 질문-대답 쌍을 포함하고 있고, 하나의 질문은 하나 또는 여러 개의 대답들을 갖고 있다. 각 질문은 범주 정보를 포함하고 있어 질문 분류 평가에 이용할 수 있다. 야후!앤써 집합에 총 101개의 범주가 있고, 이들은 세 개의 계층으로 이루어진 트리 형태이다. 본 논문에서는 범주의 크기를 생각하여 두 번째 계층의 69개 범주를 이용하였다.

분류에서 어휘 불일치 문제를 확인하고 번역확률의 유효성을 보기 위해 학습문서의 개수를 적게 구성하였다. 전체 집합의 질문 부분에서 학습 집합을 시작부분에서부터 10%로 정하였고, 파라미터 등을 결정하기 위한 측정 집합을 전체 집합의 30%되는 부분에서 10%로, 테스트 집합을 절반부분에서 시작하여 마지막까지인 50%로 정하였다. 실험에 사용되는 질문과 대답은 포터 스태머(Porter Stemmer)를 이용하여 어근처리를 하였고, 불용어(stop words)를 제거하였다.

<표 2>는 실험집합에 대한 통계정보이다. 질문에 포함된 평균 어휘 개수는 6이고 질문에 대한 대답은 어휘 번역확률을 계산할 때 사용된다.

<표 2> 야후!앤써 실험집합의 통계정보

	질의			대답
	개수	어휘 수	유일한 어휘 수	개수
전체집합	216,563	947,768	47,641	1,982,006
훈련집합	15,378	101,742	14,874	71,757
측정집합	14,259	87,904	9,935	-
테스트집합	76,259	758,122	41,492	-

5.2 비교 실험 방법

어휘 연관 정보를 반영한 분류모델이 유효하다는 것을 증명하기 위해 다음과 같이 비교실험을 하였다.

- kNN 분류기(kNN) : 최근접 이웃 분류기(k-Nearest Neighbor classifier)는 tf idf로 어휘들에 대한 가중치를 계산하였고, 코사인 유사도로 각 질문들 사이의 유사도를 계산하였다. 분류할 때 범주들의 문서 수가 같고 평균 코사인 유사도가 모두 0.7보다 크면 두 개 이상의 범주에 분류한다.
- 언어모델을 이용한 분류기(LM_k) : 언어모델로 검색한 상위 결과 k개를 이용해 분류를 한다. 스무딩(smoothing)은 Jelinek-Mercer[12] 방법으로 했고, 측정집합을 이용하여 파라미터 λ를 학습하였다.
- 번역기반 언어모델을 이용한 분류기($TransLM_k$): 어휘 번역확률을 반영한 번역기반 언어모델로 검색한 상위 결과 k개를 이용해서 분류를 한다. 측정집합을 이용하여 파라미터 β와 λ를 학습하였다.

LM_k과 $TransLM_k$는 분류할 때 범주들의 문서 수가 같으면 평균 유사도가 큰 범주로 분류하고 만일 평균 유사도가 같으면 해당 범주들에 모두 분류했다.

5.3 실험결과

비교실험 방법에 의한 성능평가는 마이크로 평균 F_1(micro-averaged F_1)을 이용하였다.

5.3.1 질문-대답 쌍의 실험결과

질문-대답 쌍들을 이용해서 어휘 번역확률을 계산 할 때 하나의 질문에 대한 각 대답을 번역 쌍으로 구성할 수 있다. 어휘 번역확률에 사용된 학습집합의 질문-대답 쌍은 71,757이며, 질문에 대한 평균 대답 개수는 4.66이다.

<표 3>은 모든 분류기에서 전체 어휘를 자질로 하였을 때, 질문-대답 쌍을 이용한 번역확률 계산에서 네 가지 번역확률 계산방법을 이용하였을 때 측정집합에 대한 성능이다. $P(QA)_{lin}$의 가중치 δ는 성능이 제일 좋을 때의 0.8을 사용하였다. 실험에 사용되는 세 분류기에서 k는 20으로 설정하였다. $TransLM_k$에서 질문-대답 쌍의 번역확률 부분에서 β가 0.9일 때 성능이 제일 좋았다. β가 높은 값을 취할 때 성능이 높은 원인은 학습집합에 어휘 수가 적어 비슷한 질문이더라도 서로 다른 어휘를 사용하여 어휘 불일치 문제가 발생하기 때문에 번역관계가 있는 다른 어휘를 찾으면 질문들이 같은 범주에 속할 확률이 높기 때문이다. $TransLM_k$

〈표 3〉 질문-대답 쌍 실험결과 (측정집합)

비교모델	번역확률 정보	micro-averaged F_1
kNN	-	0.4548
LM_k	-	0.4474
$TransLM_k$	$P(Q\|A)$	0.4651
	$P(A\|Q)$	0.4665
	$P(QA)_{lin}$	0.4624
	$P(QA)_{pool}$	0.4697

kNN과 LM_k보다 성능이 향상되었고 번역확률이 $P(QA)_{pool}$로 하였을 때 성능이 다른 번역확률보다 더 좋다.

5.3.2 질문-질문 쌍의 실험결과

각 범주에 속하는 질문-질문 쌍을 이용하여 범주에 속한 임의의 질문 Q에 대해서 언어모델로 제일 유사한 n개의 질문을 검색해서, 질문 Q에 대한 번역 쌍으로 표현하였다. 여기서 질문-질문 쌍을 QQ_n으로 표시한다.

<표 4>는 각 질문에 대해 몇 개의 질문을 번역 쌍으로 표현하느냐에 따른 질문-질문 번역 쌍 개수의 정보를 나타낸다. $QQ_{\geq 1}$은 질문-질문 유사도에서 적어도 하나의 같은 어휘를 포함하고 있으면 번역 쌍으로 표현해 준 것이고, QQ_{all}은 범주에 속한 각 질문에 대해서 다른 모든 질문들을 번역 쌍으로 표현한 것이다.

<표 5>는 번역확률 계산에서 질문-질문(Q-Q) 쌍에 대해서 학습할 때 Q-Q 쌍을 구성한 방법에 따른 성능변화를 보인 것이다. 모든 어휘를 자질로 사용하였고 측정집합에 대한 결과에서 $P(Q|Q)_{300}$일 때 성능이 제일 좋은 것을 볼 수 있다.

〈표 4〉 질문-질문(Q-Q) 번역 쌍 개수

	질의	QQ_{300}	$QQ_{\geq 1}$	QQ_{all}
번역 쌍수	15,378	3,774,282	1,572,166	7,387,992

〈표 5〉 질문-질문 쌍의 실험결과(측정집합)

	자질 / 번역 쌍	micro averaged F_1
LM_k	-	0.4474
$TransLM_k$	$P(QA)_{pool}$	0.4697
	$P(Q\|Q)_{100}$	0.4857
	$P(Q\|Q)_{200}$	0.4856
	$P(Q\|Q)_{300}$	0.4864
	$P(Q\|Q)_{500}$	0.4849
	$P(Q\|Q)_{\geq 1}$	0.4778
	$P(Q\|Q)_{all}$	0.4842

5.3.3 자질 추출후의 실험결과

분류에 영향을 미치는 중요한 자질을 선택하기 위해서 각 범주에 속하는 질문들을 이용해서 문서 빈도수(document frequency)와 카이제곱(chi square)을 이용하여 자질을 추출하였다. 여기서 각 질문 문장을 하나의 문서로 보았다. 각 범주 별로 계산한 카이제곱 값 중에서 각 어휘에 대해 가장 큰 것을 선택하여 그 어휘의 카이제곱 값으로 정하였다. 이를 통해 여러 범주에 자주 나타나는 어휘들을 제거할 수 있다. <표 6>은 야후!앤써 학습집합을 각 범주 별로 카이제곱 값으로 추출한 어휘들의 예이다.

〈표 6〉 자질 추출을 위해 범주에 대해 계산한 카이제곱 값 예

Basketball		Diet&Fitness		Mathematics		Philosophy	
basket ball	3,593.9	weight	2,164.3	x	1,031.0	life	852.0
NBA	2,419.7	diet	1,415.6	math	1,010.0	god	392.6
Kobe	471.8	lose	1,005.5	solve	419.0	philoso phy	383.4
player	396.7	fat	682.4	calculus	324.4	happy	242.3
Jordan	365.6	eat	527.1	2x	310.9	exist	232.1

〈표 7〉 자질 추출에 따른 성능 비교 (테스트 집합)

비교 모델	자질 / 번역 쌍	모든 어휘 (14,874)		df≥2인 어휘 (5,953)		χ^2≥10인 어휘 (14,608)	
LM_k	-	0.4449	-	0.4396	-	0.4466	-
$Trans\ LM_k$	$P(QA)_{pool}$	0.4669	+4.9%	0.4599	+4.6%	0.4675	+4.7%
	$P(Q\|Q)_{300}$	0.4876	+9.6%	0.4793	+9.0%	0.4881	+9.3%
	$P(Q\|Q)_{all}$	0.4850	+9.0%	0.4742	+7.9%	0.4852	+8.6%

본 실험에서의 학습문서는 크기가 작고 하나의 문서가 짧은 질문 문장으로 구성되어있어 문서 빈도수가 작은 값을 갖는 어휘들의 많다. 따라서 문서 빈도수를 기준으로 자질을 선택하였을 때와 카이제곱을 기준으로 자질을 선택하였을 때의 성능을 비교한다.

<표 7>은 자질 추출에 따른 성능변화이다. 테스트집합에 대한 실험결과로 파라미터는 측정집합에서 제일 좋은 성능을 보인 것으로 학습했다. LM_k에서 λ를 0.9로, $TransLM_k$에서 λ를 각각 $P(QA)_{pool}$일 때 0.8로, $P(Q|Q)_{300}$일 때 0.7로, $P(Q|Q)_{all}$일 때 0.6으로 하였고, 세 분류기의 β를 모두 0.9로 하였다. 질문-질문 쌍의 번역확률 계산에서 측정집합에서 좋은 성능을 보인 $P(Q|Q)_{300}$에서 테스트 집합에서도 가장 좋은 성능을 보였다. 자질 추출에서 카이제곱을 이용했을 때 성능이 문서 빈도수를 이용했을 때보다 향상되었다. 카이제곱으로 자질 추출 후 $TransLM_k$분류방법이 LM_k보다 성능이 $P(QA)_{pool}$일 때 4.7%, $P(Q|Q)_{300}$일 때 9.3% 향상되었다. $TransLM_k$에서 Q-Q쌍으로 계산한 번역확률을 적용한 후 성능이 Q-A쌍보다 더욱 좋다.

5.4 결과 분석

$TransLM_k$의 성능이 언제나 LM_k보다 좋은 것은 학습문서의 질문과 분류하려는 질문 사이에 같은 어휘가 존재하지 않더라도 번역확률에 의해 학습집합에서 적절한 질문을 찾아주기 때문이다. LM_k방법과 $TransLM_k$방법에서 아래 질문에 대한 검색결과 k를 5로 했을 때 상위 검색 결과는 <표 8>, <표 9>와 같다.

- 질문 "I NEED TO LOOSE 40 POUNDS BY NEXT SUNDAY! help please?"

질문의 정답 범주는 'Diet&Fitness'이다. LM_k분류방법은 이 질문을 'Football (American)' 범주로 잘못 분류하지만, $TransLM_k$분류방법은 이 질문을 'Diet&Fitness'로 정확하게 분류하였다.

〈표 8〉 LM로 검색한 상위 5개 질문

순위	학습문서에 검색된 질문	범주
1	What television network if any, will carry Sunday Night Football next season?	Football (American)
2	need help.. please.^-^?	Botany
3	need help please?	Homework Help
4	I NEED HELP PLEASe?	Mathematics
5	Loosing fat?	Diet&Fitness

〈표 9〉 TransLM로 검색한 상위 5개 질문

순위	학습문서에 검색된 질문	범주
1	I want to **lose** 20 **pounds** easily any suggestion?	Diet&Fitness
2	Guide to **loose weight**?	Diet&Fitness
3	**Loseing weight**...i **need help**?	Diet&Fitness
4	is 16 ounces one **pound**?	Mathematics
5	what should I do to **loose weight**	Diet&Fitness

질문에 나타난 어휘 'loose'와 'pound'에 대해서 $TransLM_k$에서 이용한 어휘 연관 정보는 다음과 같다. 어휘 번역확률은 QQ_{300}에서 학습한 것이다.

p(loose|weight) = 0.0099 p(pound|lose) = 0.0080 p(loose|lose) = 0.0062 p(pound|loose) = 0.0024 p(loose|pound) = 0.0024 p(pound|weight)=0.0065

6. 결 론

본 논문에서는 같은 범주에 속하는 질문-질문 쌍들에 대해서 번역확률을 계산하여 번역기반 언어모델로 질문을 분류하는 방법을 제안하였다.

질문 분류에서 어휘 연관성을 반영한 것이 학습집합의 크기가 작고 질문이 짧은 상황에서 유용하였다. 카이제곱으로 자질 추출 후 $TransLM_k$분류방법이 LM_k보다 성능이 $P(QA)_{pool}$일 때 4.7%, $P(Q|Q)_{300}$일 때 9.3% 향상되었다. $TransLM_k$의 성능이 언제나 LM_k보다 좋은 것은 학습문서의 질문과 분류하려는 질문 사이에 같은 어휘가 존재하지 않더라도 어휘 번역확률에 의해 학습집합에서 적절한 질문을 찾아주기 때문이다. 야후!앤써 집합으로 실험한 결과로 보면 범주 내에서의 질문-질문 쌍으로 계산한 번역확률이 질문-대답 쌍으로 계산한 번역확률보다 분류 성능이 더 우수하다는 것을 볼 수 있다.

참 고 문 헌

[1] KDDCUP 2005, http://www.acm.org/sigs/kddcup/

[2] Yangdong Liu, Jiang Bian and Eugene Agichtein, "Predicting Information Seeker Satisfaction in Community Question Answering," Proceeding of the 31st Annual International ACM SIGIR Conference, pp.483-490, July, 2008.

[3] A. Berger and J. Lafferty, "Information retrieval as statistical translation," Proceedings of the 22nd annual international ACM SIGIR conference, pp.222-229, Aug., 1999.

[4] Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee, "Finding Similar Questions in Large Question and Answer Archives," Proceedings of the 14th ACM SIGIR Conference, pp.84-90, 2005.

[5] Xiaobing Xue, Jiwoon Jeon and W. Bruce Croft, "Retrieval Models for Question and Answer Archives," Proceedings of the 31st annual international ACM SIGIR conference, pp.475-482, July, 2008.

[6] Huanhuan Cao, Derek HaoHu, Dou Shen and Daxin Jiang, "Context-Aware Query Classification," Proceedings of the 32nd annual international ACM SIGIR conference, pp.3-10, July, 2009.

[7] Dou Shen, Jian-Tao Sun, Qiang Yang and Zheng Chen, "Building Bridges for Web Query Classification," Proceedings of the 29th annual international ACM SIGIR conference, pp.131-138, Aug., 2006.

[8] ODP, http://dmoz.org

[9] Yu Jingbo and YeNa, "Automatic Web Query Classification Using Large Unlabeled Web Pages", Proceedings of the 2008 The Ninth International Conference, pp.211-215, 2008.

[10] Dell Zhang, Wee Sun Lee, "Question Classification using Support Vector Machines", Proceedings of the 26th annual international ACM SIGIR conference, pp.26-32, 2003.

[11] GIZA tool, http://www.fjoch.com/GIZA++.html

[12] ChengXiang Zhai, John Lafferty, "A study of smoothing methods for language models applied to information retrieval", ACM Trans.Inf.Syst, Vol.22, No.2, pp.179-214, 2004.

[13] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics 19, 2(1993), pp.263-311.

김 설 영

e-mail : xying@chonbuk.ac.kr

2008년 장안대학교 컴퓨터공학과 학사

2008년~현 재 전북대학교 컴퓨터공학과 석사과정

관심분야 : 정보검색, 정보 마이닝, 자연언어처리

이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과 학사

1997년 한국과학기술원 전자전산학 석사

2001년 한국과학기술원 전자전산학 박사

2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현 재 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수

관심분야 : 정보검색, 정보 마이닝, 자연언어처리