유사한 질의쌍의 어휘 번역확률을 이용한 질의 분류

김설영, 장계훈, 이경순
전북대학교 컴퓨터공학과
e-mail : xying@chonbuk.ac.kr, ghjang@chonbuk.ac.kr, selfsolee@chonbuk.ac.kr

Query Classification Based on Translation Probabilities of Similar Query Pair

Xueying Jin, Kye-Hun Jang, Kyung-Soon Lee
Dept. of Computer Engineering, Chonbuk National University

요 약

질의 분류에서 어휘의 다양한 표현으로 인한 어휘 불일치문제는 성능저하의 주요 원인이다. 본 논문에서는 야후!앤써 질의응답 아카이브를 이용해서 같은 카테고리의 질의-질의쌍들에 대해 어휘-어휘 번역확률을 계산하는 방법을 제안한다. 정보검색에서 우수한 성능을 보인 어휘 사이의 번역확률을 반영하는 번역기반 언어모델이 질의 분류에서 유효함을 확인하였고 언어모델과의 비교실험을 통해 성능향상을 보였다. 어휘관계를 측정하는 방법에서 번역확률 계산방법에 따른 성능측정에서 전체 질의-대답쌍들에 대해 번역확률을 계산하는 것보다 같은 카테고리에 속하는 질의-질의쌍들에 대해 번역확률을 계산하는 것이 분류를 위해 더 좋은 번역확률임을 확인하였다.

1. 서론

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

본 논문에서는 질의 분류의 어휘 불일치 문제를 해결하기 위해 어휘 관계(word relationships) 정보를 반영한 번역모델을 이용하여 질의를 분류한다. 야후!앤써(Yahoo!Answers)[2] 컬렉션의 학습집합에 대해서 같은 카테고리에 속하는 질의-질의(Q-Q)쌍을 이용해 번역확률을 계산하여 어휘관계정보를 획득하는 방법을 제안한다. 같은 카테고리내의 질의들은 그 카테고리를 대표하는 주제를 표현할 것이다.

제안한 방법의 유효성을 검증하기 위해 언어모델을 이용한 분류기와 비교실험을 하였다. 질의 분류에서 번역확률 계산 방법에 따른 성능 변화를 관찰하기 위하여 어휘들 사이의 번역확률을 질의-대답(Q-A)쌍을 이용해서 계산하는 방법과 각 카테고리에서 질의-질의쌍을 이용하여 번역확률을 계산하는 방법을 비교하였다. 질의 분류에서 카테고리 정보를 이용하여 카테고리 내에서 질의-질의쌍으로 계산한 번역확률이 질의-대답쌍으로 계산한 번역확률보다 분류 성능이 더 우수하다.

본 논문의 구성은 다음과 같다. 2 장에서 관련 연구에 대해 기술하고, 3 장에서는 어휘 번역확률을 이용한 질의 분류에 대해 소개하고, 4 장에서는 어휘관계 정보 획득에 대해 설명하고, 5 장에서는 실험 및 평가, 6 장에서는 결론 및 향후 연구에 대해 언급한다.

2. 관련 연구

정보검색에서 Berger 과 Lafferty[3]는 단어들 사이의 번역 확률을 이용하여 IBM 모델 1 로 문서들을 검색하는 방법을 소개하였다. Jiwoon[4]는 IBM 모델과 언어모델을 결합한 번역기반 언어모델을 제안하였다. 질의응답 아카이브에서 어휘 번역확률을 계산하여 질의응답 검색을 하여 번역기반 언어모델이 언어모델에 비해 성능이 향상됨을 보였다. Xiaobing[5]은 [4]에서 제안한 번역기반 언어모델에서 번역확률을 계산하는 방법에 따른 성능 변화를 비교하였다. 또한 질의응답 검색에서 질의 문장뿐만 아니라 대답 문장에 대한 언어모델 검색결과를 반영하였다. 번역확률 계산 방법은 질의와 대답을 쌍으로 묶어주는데 있어서, 질의-대답 쌍 또는 대답-질의쌍 등 질의와 대답을 각각 번역 부분의 소스로 하였을 때의 확률을 계산하는 방법과 질의-대답쌍과 대답-질의쌍을 하나의 풀(pool)에

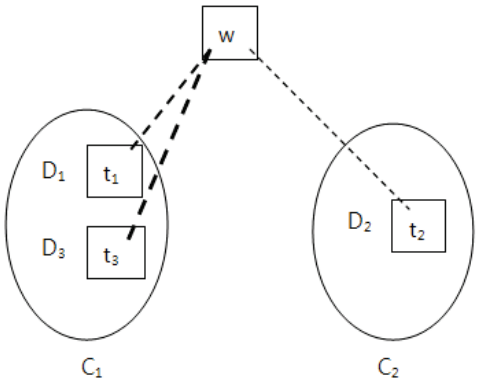
넣고 전체에 대해 번역확률을 계산하는 방법 등에 대해 비공개 질의응답 검색 실험집합인 원더 컬렉션(Wondir Collection)에 대해 실험하였다. 본 논문에서는 질의 분류를 위하여 같은 카테고리의 질의-질의쌍들에 대해 번역확률을 계산하였다.

질의 분류는 이미 정해진 카테고리들에 대해서 새로운 질의를 하나 이상의 카테고리로 분류하는 것이다. Cao[6]는 사용자가 검색엔진에 연속 올려놓은 인접한 질의들은 의미상 연관된 질의들이란 가정을 이용하여 질의 A 의 카테고리 정보를 이용하여 사용자의 의향을 파악하여 질의 B 의 카테고리를 판단한다. 이와 같이 문맥 정보를 이용하는 방법은 온라인 모드의 실제 분류 시스템에 큰 도움이 된다. Dou Shen[7]는 KDDCUP 2005 데이터 집합의 질의와 카테고리를 확장(enrich)하여 분류하였다. 질의와 목표 분류 체계 사이에 중간 분류 체계를 삽입하여 질의를 먼저 중간 카테고리에 분류한 후 중간 분류 체계와 목표 분류 체계가 갖고 있는 직접 매칭 혹은 확장 매칭에 의해 질의를 목표 카테고리로 분류하였다. 이 방법은 온라인 모드에서 목표 분류 체계의 카테고리가 변해도 질의를 중간 분류 체계에 분류할 수 있어 최종적으로 질의를 목표 카테고리로 분류한다.

3. 어휘 번역확률을 이용한 질의 분류

본 논문에서 질의 분류는 검색을 통한 상위의 검색결과를 이용해서 검색 문서가 많이 속하는 카테고리로 분류를 하였다. 어휘관계 정보를 반영한 분류 형태는 (그림 1)에 나타나있다. 카테고리 C_1 과 C_2 의 문서들에 단어 w 가 존재하지 않기에 단어매칭으로 문서들을 검색할 수 없다. 어휘-어휘 관계를 이용하여 질의에 어휘 w 가 존재하지 않지만 w 와 어휘관계가 있는 어휘 t_i(i=1,2,3)가 있을 때 모델은 어휘 w 와 t_i 의 관계, 번역확률을 이용해서 t_i 를 포함한 문서들을 검색해 준다. 문서의 검색 순위는 w 와 t_i 사이의 번역확률과 관계된다. 점선의 굵기가 어휘들 사이의 번역확률 크기를 나타날 때 w 는 t_3 과의 확률이 제일 크고 그 다음으로 t_1 과 t_2 이다. 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 학습 문서들 중에서 질의와 유사도가 가장 높은 순위인 k 개의 문서를 구하고 그들이 가장 많이 속한 카테고리로 분류한다. 문서 D_1 과 D_3 는 카테고리 C_1 에 속하고 D_2 는 카테고리 C_2 에 속한다. k=3 일 때 C_1 에 속하는 문서가 두 개 있고 C_2 에 속하는 문서가 한 개 있으므로 질의 Q 는 C_1 로 분류하게 된다.

어휘관계 정보를 반영하는 정보검색 모델인 번역기반 언어모델(Translation-based Language Model)은 정보검색 모델의 하나인 언어모델(Language Model) 과 어휘 번역확률을 반영하는 IBM 모델 1[3]을 개선한 방법으로,질의응답 아카이브에 대한 검색과 일반 정보검 정보검색에서 언어모델에 비해 우수한 성능을 보



(그림 1) 어휘 관계 정보를 이용한 분류 예

였다[4].

번역기반 언어모델을 질의분류에 적용시키기 위해서는 어휘에 대한 번역확률을 획득하는 방법이 필요하다.

번역기반 언어모델의 기본이 되는 언어모델의 수식은 다음과 같다.

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

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

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

Q 는 질의를 나타내고, D 는 검색할 문서를 나타낸다. |D|는 문서의 길이로 문서 D 에 나타난 어휘의 개수를 나타낸다. 번역기반 언어모델은 어휘 번역확률 정보를 언어모델에 반영시키기 위해 수식 (2)를 아래 수식 (4)와 (5)로 변형하였다.

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

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

P(w|t) 는 어휘 t 가 w 로 번역된 확률이고 $P_{ml}(t|D)$는 어휘 t 에 대한 그 문서에서의 확률이다.

4. 어휘 관계 정보 획득

일반적으로 어휘-어휘 번역확률은 기계번역(Machine Translation)에서 서로 다른 언어 쌍에 대해서, 예를 들어 한국어 문장을 소스(source)로 보고, 이에 대한 영어 번역문장을 타깃(target)으로 보고, 한국어 어휘에 대한 영어 어휘의 번역확률을 계산한다.

본 논문에서 어휘사이의 번역확률을 계산하는 방법으로 질의-대답쌍을 이용하는 방법과 질의 분류를 위한 각 카테고리에 속하는 질의-질의쌍을 이용한 방법으로 어휘 번역확률을 계산하였다. 어휘 번역확률은 EM 알고리즘을 이용한

GIZA++[8]를 이용하여 계산하였다.

4.1 질의-대답쌍을 이용한 어휘 번역확률 계산

질의 $q_1,q_2,\ldots,q_M$ 으로 구성된 질의 집합을 $Q=\{q_1,q_2,\ldots,q_M\}$라 하고, 대답 $a_1,a_2,\ldots,a_N$ 으로 구성된 대답 집합을 $A=\{a_1,a_2,\ldots,a_N\}$라 할 때, 질의와 대답으로 구성된 질의-대답 쌍 $(q,a)_i$ 는 질의를 소스로, 대답을 타깃으로 본 것이다.

$P(w_i|w_j)$는 소스 어휘가 w_j 일 때 타깃 어휘가 w_i 일 번역확률을 나타낸다. $P(w_i,A|w_j,Q)$는 질의(Q)를 소스, 대답(A)를 타깃으로 하였을 때의 번역확률이다. 여기서부터 P(A|Q)로 표현한다. P(A|Q)는 질의-대답 쌍 컬렉션$\{(q,a)_1,\ldots,(q,a)_n\}$으로 학습한다.

$P(w_i,Q|w_j,A)$는 대답(A)를 소스, 질의(Q)를 타깃으로 하였을 때의 번역확률이며 P(Q|A)로 하며 대답-질의쌍 컬렉션 $\{(a,q)_1,\ldots,(a,q)_n\}$으로 학습한다.

P(A|Q)와 P(Q|A)를 결합하는데 두 가지 방법이 있다. 첫 번째 방법은 질의-대답쌍과 대답-질의쌍을 컬렉션 $\{(q,a)_1,\ldots,(q,a)_n,(a,q)_1,\ldots,(a,q)_n\}$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 수식으로 표현하면 아래와 같다.

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

4.2 각 카테고리에 속하는 질의-질의쌍을 이용한 어휘 번역확률 계산

질의 분류에서 같은 카테고리에 속하는 질의들은 카테고리를 대표하는 어휘들이 포함되어 있으므로 어휘 사이의 관계는 분류를 위한 중요한 정보가 될 수 있다. 따라서 각 카테고리에 속하는 질의-질의를 번역 쌍으로 구성하여 번역확률을 계산하였다.

$P(w_i,Q_p|w_j,Q_q)$는 Q_q 를 소스로, Q_p 를 타깃으로 $(p\neq q)$로 하였을 때의 번역확률이다. 어떤 한 카테고리에 속하는 질의들에 대해서 하나의 질의 Q_q 에 임의의 개수의 질의 Q_p 를 번역 쌍으로 만들 수 있다. 모든 카테고리의 번역 쌍들을 하나의 풀에 넣고 번역확률을 구한다. $P(Q|Q)_n$ 은 n 개의 Q-Q 쌍을 만들었을 때의 확률이다.

<표 1>은 질의-대답 쌍과 대답-질의 쌍을 이용한 'lose'에 대한 번역확률의 일부 예이다.

<표 1> 어휘 'lose'에 대한 번역확률 일부 예

질의-대답 쌍		질의-질의 쌍			
$P(QA)_{pool}$		$P(Q\|Q)_{100}$		$P(Q\|Q)_{300}$	
lose	0.1211	lose	0.1726	weight	0.1048
eat	0.0480	weight	0.1564	lose	0.0788
exercise	0.0376	fat	0.0182	diet	0.0180
weight	0.0375	diet	0.0180	loss	0.0151
lost	0.0351	way	0.0138	fat	0.0149

5. 실험 및 분석

5.1 야후!앤써 실험집합

어휘관계 정보를 반영한 질의 분류의 유효성을 평가하기 위해 야후!앤써 실험집합[2]을 이용하였다. 야후!앤써 집합에 총 101 개의 카테고리가 있고, 이들은 세 개의 계층으로 이루어진 트리 형태이다. 본 논문에서는 두 번지 계층의 69 개 카테고리를 이용하였다.

분류에서 어휘 불일치 문제를 확인하고 어휘 관계 정보의 유효성을 보기 위해 학습문서의 개수를 적게 구성하였다. 전체 집합의 질의 부분에서 학습 집합을 10%로 정하였고, 파라미터 등을 결정하기 위한 측정 집합을 10%로, 테스트 집합을 50%로 정하였다. 실험에 사용되는 질의와 대답은 포터 스태머(Porter Stemmer)를 이용하여 어근처리를 하였고, 불용어(stop words)를 제거하였다.

<표 2>는 실험집합에 대한 통계정보이다. 질의에 포함된 평균 어휘 개수는 6 이다.

<표 2> 야후!앤써 실험집합 통계정보

	질의 개수	질의의 어휘 개수	질의의 유일한 어휘개수	대답 개수
전체집합	216,563	947,768	47,641	1,982,006
훈련집합(10%)	15,378	101,742	14,874	71,757
측정집합(10%)	14,583	87,904	9,935	-
테스트집합(50%)	76,259	758,122	41,492	-

5.2 비교실험 방법

실험에 사용될 분류기는 kNN 분류기(kNN), 언어모델을 이용한 분류기(LM_k), 및 번역기반 언어모델을 이용한 분류기($TransLM_k$)가 있다. kNN 분류기에서 tf, idf 로 어휘들의 가중치를 구하고 코사인 유사도로 질의들 사이의 유사도를 계산하였다. LM_k 와 $TransLM_k$ 에서는 LM 모델과 TransLM 모델로 각각 검색한 상위 k 개 결과를 이용해서 분류하고, 스무딩(smoothing)은 Jelinek-Mercer 방법으로 측정집합에서 파라미터 λ를 학습하였다.

5.3 실험결과

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

질의-질의쌍의 번역확률을 계산하기 위해 카테고리에 속한 임의의 질의 Q 에 대해서 언어모델로 제일 유사한 n 개의 질의를 검색해서, 질의 Q 에 대한 번역쌍을 구성하여 확률 $P(Q|Q)_n$ 을 계산한다. $P(Q|Q)_{\geq 1}$ 은 질의-질의 유사도에서 적어도 하나의 같은 어휘를 포함하면 번역쌍을 구성하여 번역확률을 계산하고,

<표 3> 번역확률 계산방법에 따른 측정집합 실험결과(측정집합)

비교모델	자질 / 번역 쌍	micro averaged F_1
kNN	-	0.4548
LM_k	-	0.4474
$TransLM_k$	$P(Q\|A)$	0.4651
	$P(A\|Q)$	0.4665
	$P(QA)_{lin}$	0.4621
	$P(QA)_{pool}$	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

$P(Q|Q)_{all}$ 은 카테고리의 각 질의에 대해서 다른 모든 질의들을 번역쌍으로 표현하여 계산한 번역확률이다.

<표 3>는 모든 어휘를 자질로 사용하였을 때 4 가지 질의-대답쌍 번역확률과 질의-질의쌍 번역확률을 이용한 측정집합에 대한 성능이다. 모든 분류기에서 k 는 20 으로 설정하였고 TransLMk 의 β 가 0.9 일 때 성능이 가장 좋다. $TransLM_k$ 의 성능이 kNN 과 LM_k 보다 향상되었고, 4 가지 질의-대답 번역확률에서 $P(QA)_{pool}$ 의 성능이 기타 번역확률보다 더 좋았다. 질의-질의 번역확률에서 $P(Q|Q)_{300}$ 일 때 성능이 제일 좋았고, $P(QA)_{pool}$ 보다 성능이 더욱 향상되었다.

분류에 영향을 미치는 중요한 자질을 선택하기 위해 카이제곱을 이용하였다. 카이제곱이 10 이상인 유일한 어휘는 14,608 개이다. 각 카테고리 별로 계산한 카이제곱 값 중에서 각 어휘에 대해 가장 큰 것을 선택하여 그 어휘의 카이제곱 값으로 정하였다.

<표 4>은 자질 추출에 따른 테스트집합의 성능결과이다. λ 는 각각 LM_k 에서 0.9 로, $TransLM_k$ 의 $P(QA)_{pool}$ 일 때 0.8 로, $P(Q|Q)_{300}$ 일 때 0.7 로, $P(Q|Q)_{all}$ 일 때 0.6 으로 하였으며, 세 분류기의 β 를 모두 0.9 로 하였다. $TransLM_k$ 에서 Q-A 쌍으로 계산한 번역확률을 적용한 성능보다 Q-Q 쌍으로 계산한 번역확률을 적용한 후 성능이 더욱 좋다.

$TransLM_k$ 의 성능이 언제나 LM_k 보다 좋은 것은 학습문서의 질의와 분류하려는 질의 사이에 같은 어휘가 존재하지 않더라도 어휘관계에 의해 학습집합에서 적절한 질의를 찾아주기 때문이다.

<표 4> 자질 추출에 따른 성능비교(테스트집합)

비교모델	자질 / 번역쌍	모든 어휘 (14,874)	**χ^2>10 인 어휘 (14,608)**
LM_k	-	0.4449	0.4466
$TransLM_k$	$P(QA)_{pool}$	0.4669(+4.9%)	0.4675(+4.7%)
	$P(Q\|Q)_{300}$	0.4876(+9.6%)	**0.4881(+9.3%)**
	$P(Q\|Q)_{all}$	0.4850(+9.0%)	0.4852(+8.6%)

6. 결론 및 향후 연구

본 논문에서는 번역확률 계산방법에서 전체 질의-대답쌍들에 대해서 번역확률을 계산하는 것보다 같은 카테고리에 속하는 질의-질의쌍들에 대해서 번역확률을 계산하는 것이 분류를 위해서 더 좋은 번역확률임을 확인하였다. 질의분류에서 어휘관계 정보를 반영한 것이 학습집합의 크기가 작고 질의가 짧은 상황에서 유용하다는 것을 확인하였다. 실험에서 질의-대답쌍을 이용한 번역확률을 학습한 $TransLM_k$ 분류방법이 보다 LM_k 보다 성능이 모든 어휘일 때 4.9% 향상되었고, 카이제곱을 이용한 자질 추출 후 4.7% 향상되었다. 각 카테고리에 속하는 질의-질의쌍을 이용한 번역확률 계산방법이 질의-대답쌍을 이용한 번역확률 계산방법보다 질의 분류에서 우수함을 확인하였다. $TransLM_k$ 는 번역확률이 $P(Q|Q)_{300}$ 이고 자질을 추출하였을 때 LM_k 보다 9.3% 향상되었다. 실험결과에서 보면 카테고리 내에서의 질의-질의쌍으로 계산한 번역확률이 정보검색에서 사용된 질의-대답쌍으로 계산한 번역확률보다 분류성능이 더 우수하다.

향후 연구에서 어휘들의 의존관계(term dependency)를 이용하여 질의를 분류하려 한다. 질의에 늘 함께 나타나는 어휘들 사이에는 어떤 의존관계가 있다고 추측할 수 있다. 이런 의존된 어휘들을 이용하면 분류 성능을 더욱 향상시킬 것이다.

참고문헌

[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" Proceedings 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 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. "Concext-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] GIZA tool. http://www.fjoch.com/GIZA++.html