어휘관계 정보와 질의개념연관도를 반영한 정보검색 성능 향상 기법

김준길, 이경순
전북대학교 컴퓨터공학과
e-mail : {jgkim, selfsolee}@chonbuk.ac.kr

Information Retrieval Based on Word Relationships and Degree of Query Concept

Jun-Gil Kim, Kyung-Soon Lee
Dept. of Computer Engineering, Chonbuk National University

요 약

정보검색 분야에서 어휘 불일치 문제를 해결하기 위해 질의에서의 어휘 사이의 관계를 반영하는 것은 필수적인 요구사항이 되었다. 본 논문에서는 문장-문장 번역쌍을 이용하여 어휘 번역확률을 계산하였고, 어휘관계 정보를 반영하는 번역기반 언어모델에 어휘와 질의 개념과의 연관 정도를 반영한 모델을 제안한다. 뉴스 컬렉션 집합인 TREC AP 컬렉션에 대한 비교실험을 하였다. 실험결과에서 언어모델보다 어휘 관계를 반영한 번역기반 언어모델의 성능이 향상되었고 어휘의 질의개념 연관도를 반영한 모델이 번역기반 언어모델보다 성능이 향상됨을 보였다.

1. 서론

정보검색 분야에서 어휘의 다양한 표현으로 인한 어휘 불일치 문제(word mismatch problem)는 검색 성능 저하의 주요 원인이다. 이러한 어휘 불일치 문제를 해결하기 위하여 어휘관계정보(word relationships)를 반영한 정보검색 모델에 관한 연구[1,2,3,4]가 활발히 이루어지고 있다. 어휘관계정보를 반영함으로써 질의가 문서에 나타나지 않았더라도 질의와 연관된 어휘가 포함된 문서를 검색하는 것이 가능하다.

어휘관계정보를 반영한 기존연구로 Berger & Lafferty[1]는 어휘들 사이의 번역 확률을 이용하여 IBM 모델 1[2] 로 문서들을 검색하였다. Murdock[3]은 문장 검색(Sentence Retrieval)에서의 번역모델(Translation Model)을 제안하였고, Jeon[4]은 질의응답 아카이브에서 유사한 질문을 찾는 문제에서 IBM 모델 1 과 언어모델(Language Model)을 결합한 번역기반 언어모델(Translation-based Language Model)을 제안하였다. 질의들 사이에 같은 어휘가 존재하지 않더라도 어휘관계 정보를 이용하여 비슷한 질의를 찾아주었다.

번역기반언어모델에서 같은 단어 사이의 어휘번역 확률을 적용할 때 자기번역 문제(self-translation problem)가 발생한다. 자기번역 문제란 어떤 어휘가 자기 자신으로의 변역확률 값을 가지게 되는데 자신의 번역확률 값이 낮을 경우에는 매칭되는 어휘에 대해서 낮은 검색 성능을 보여주게 된다. 이와 반대로 자신의 번역확률 값이 매우 높을 경우에는 어휘관계 정보를 이용하는 가치가 없어지게 된다[7]. 이러한 문제들을 해결하기 위하여 Jeon[5]은 모든 어휘에 대하여 자신의 번역확률을 1 로 고정했다. Jin[6]은 자신의 번역확률을 자신 이외의 모든 어휘로의 번역확률 값의 합보다 항상 크거나 같게 설정하여 실험했다.

질의응답 아카이브에서 Xue[7]은 Jiwoon[4] 이 제안한 번역기반 언어모델에 번역확률 적용방법을 제안하여 실험하였다. 질의와 대답을 쌍으로 묶어주어 질의와 대답을 각각 번역부분의 소스로 하였을 때의 확률계산방법과 질의-대답쌍과 대답-질의쌍을 결합하여 확률을 계산하는 방법을 번역기반 언어모델로 비교실험하였으며 번역기반 언어모델의 유효성을 보였다.

질의응답 아카이브에서 구축한 어휘번역확률 정보를 정보검색에 이용하기에는 검색대상인 컬렉션의 성격이 다르다는 문제점이 있다.

본 논문에서는 보다 일반적인 정보검색에서의 어휘관계정보를 위하여 검색대상인 문서에 대해서 번역확률을 계산하는 방법을 제안한다. 또한, 어휘의 질의개념 연관도를 반영하여 검색 성능을 향상시키는 방법을 제안한다. 어휘의 질의개념(Query Concept) 연관도는 어휘가 질의가 내포하는 개념과 어느정도 관계를 갖는지를 반영하는 것이다.

문서에서 한 문장과 그 다음 문장 사이에는 내용의 흐름에 있어서 연관성이 있다는 것을 가정하고, 문장 사이의 어휘 번역확률을 계산하였다. 어휘관계 정보를 반영하는 번역기반 언어모델에 어휘와 질의 개념과의 연관 정도를 반영한 모델을 개발하였다. 또한, 자기번역 문제에 대해서 자신의 번역확률을 0 과 1 그리고 계산한 어휘번역확률 값을 이용하는 세가지 방법으로 비교 실험하였다. 본 논문에서 제안한 방법

의 유효성을 검증하기 위해서 TREC AP 컬렉션에 대해서 실험하였다.

2. 어휘 번역확률 정보 획득 방법

일반적으로 어휘-어휘 번역확률은 기계번역(Machine Translation)에서 서로 다른 언어 쌍에 대해서 번역 확률을 계산하는 것이다. 예를 들어, 한국어와 영어 문장을 이용하여 번역확률을 계산할 경우 한국어 문장을 소스(source)로 보게 되고 영어 문장을 타깃(target)으로 보고 번역확률을 계산하게 된다.

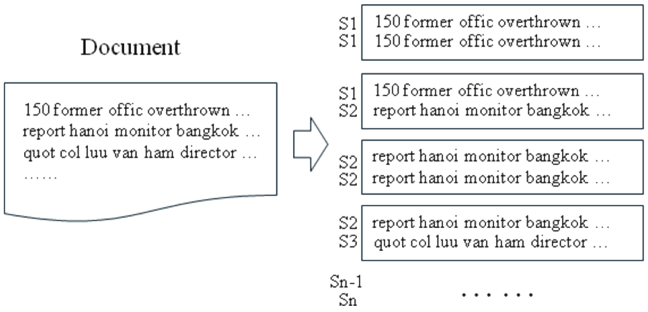
P(w|t)는 소스 어휘가 t 일 때 타깃 어휘가 w 일 번역 확률을 나타낸다. 어휘-어휘 번역확률은 두 어휘 사이에 관계가 있음을 나타내고, 그 관계의 중요도를 번역확률 값으로 나타낸다.

어휘-어휘 사이의 관계정보를 얻기 위해서 본 논문에서는 *하나의 문서 내에서 어떤 한 문장과 그 다음 문장 사이에는 내용의 흐름에 있어서 연관성을 가지고 기술되어 있다고 가정한다.* 따라서, 자기 자신의 문장의 번역쌍과, 자신의 문장과 그 다음에 나타나는 문장을 같은 언어에서의 번역 쌍으로 보고, 어휘번역확률을 계산한다. 검색대상 문서집합인 TREC AP 컬렉션에서 번역확률을 계산하였다.

문서에서 나타난 문장을 {S1, S2, …, Sn}이라 할 때, 문장 S1 과 그 다음 문장 S2 는 내용의 흐름에 있어서 연관성이 있다고 보고 S1-S2 번역쌍으로 표현해서 전체 컬렉션에 있는 번역쌍들을 이용해서 학습을 한다.

문장 쌍을 이용한 번역쌍 표현은 (그림 1)과 같이, 자신의 문장과 자신의 문장을 번역쌍으로 보고, 자신과 그 다음문장을 번역쌍으로 표현하였다. 어휘 번역확률 계산을 위한 TREC AP 컬렉션에서의 문장 번역쌍 개수의 정보는 <표 1>과 같다.

어휘-어휘 번역확률을 계산하는 간단한 방법은 두 어휘의 공기빈도(co-occurrence)를 이용하는 것이다. 만일 한 타깃 어휘가 주어진 소스 어휘와 여러 번 같이 나타났다면, 이 타깃 어휘는 소스 어휘의 번역 어휘가 될 확률이 아주 높다. 그러나 이 방법을 단순히 적용하면 대량의 불용어에 큰 확률을 주게 된다. 그것은 불용어가 소스 어휘와 자주 함께 나타나기 때문이다. 따라서 의미 없는 공동 발생을 배제하는 더욱



(그림 1) 문서에서 문장 번역쌍을 {S1-S1, S1-S2}로 표현한 예제

좋은 방법이 필요하다.

본 논문에서는 어휘 번역 확률을 EM 알고리즘을 이용한 GIZA++[8]를 이용하여 번역 확률을 계산하였다. 문장-문장 번역쌍을 이용한 어휘 번역확률 계산 예제는 <표 2>와 같다. 학습질의 'hubble space telescope'에 나타난 어휘에 대한 번역확률을 나타냈다.

3. 어휘의 질의 개념 연관도를 반영한 번역기반 언어모델

어휘관계 정보를 정보검색에 반영하기 위하여 언어모델(Language Model)과 어휘번역확률을 반영하는 IBM 모델 1[2]을 개선한 번역기반 언어모델(Translation-based Language Model)[4,7]을 사용하였다.

번역기반 언어모델의 수식은 다음과 같다.

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

$$P(w \mid D) = \frac{|D|}{|D|+\mu} P_{mx}(w \mid D) + \frac{\mu}{|D|+\mu} P_{ml}(w \mid C) \quad (2)$$

$$P_{mx}(w \mid D) = (1-\beta) P_{ml}(w \mid C) + \beta \sum_{t \in D} P(w \mid t) P_{ml}(t \mid D) \quad (3)$$

$$P_{ml}(w \mid D) = \frac{freq(w, D)}{|D|}, P_{ml}(w \mid C) = \frac{freq(w, C)}{|C|} \quad (4)$$

여기서 Q 는 질의를 나타내고 D 는 문서, C 는 컬렉션을 의미한다. |D|는 문서에 나타난 어휘의 개수이고 |C|는 컬렉션에 나타난 어휘의 총 개수가 되고, freq(w,

<표 1> 어휘 번역확률 계산을 위한 TREC AP 컬렉션에서의 문서 정보

문서 정보	문서 개수	문장 개수	문장 번역쌍 개수	유일한 단어 개수	전체 단어 개수
AP88	79,919	1,681,071	3,081,041	131,901	36,005,601
AP89	84,678	1,809,844	3,231,851	139,308	38,599,290
AP90	78,321	1,740,612	3,095,866	134,561	35,978,387

<표 2> 문장-문장 번역쌍을 이용한 번역확률 예제

	번역확률에서 소스 부분 어휘					
	hubble		space		telescope	
타깃 어휘	hubble	0.5463	space	0.6435	telescope	0.6367
	weiler	0.1185	weightless	0.1693	weiler	0.1019
	fainter	0.0823	spacewalk	0.1276	hubble	0.0700
	telescope	0.0597	akiyama	0.1188	fisk	0.0672
	danburi	0.0407	soyuz	0.1068	supernova	0.0645
	crisp	0.0337	cosmonaut	0.1047	goddard	0.0532

D)는 어휘 w 가 문서 D 에 나타낸 빈도수이고 freq(w, C)는 어휘 w 가 컬렉션 C 에 나타난 빈도수이다.

번역기반 모델에서 어휘 번역확률 정보를 반영시키는 수식(3)에서 $P(w|t)$ 는 어휘 t 가 w 로 번역되는 확률이다. 질의에 나타난 어휘 w 가 문서에 나타나지 않더라도 문서에 나타난 어휘 t 와 w 와의 관계를 이용하여 검색에 반영한다. β 값으로 어휘 번역확률 부분의 중요도를 조정할 수 있다. 만약 β 에 0 을 부여한다면 언어모델과 같아지게 된다.

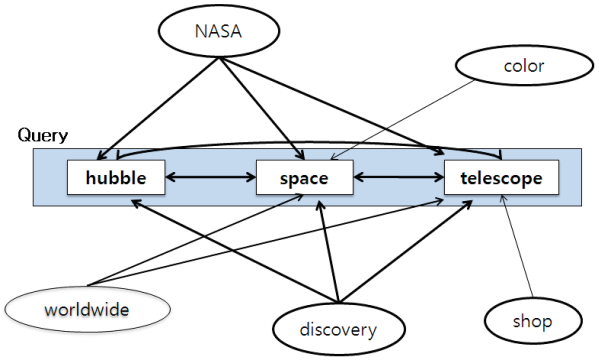
본 논문에서는 어휘관계 정보를 반영하는 번역기반 언어모델에 어휘와 질의 개념과의 연관 정도를 반영한 모델을 개발하였다. 위 수식(3)을 다음과 같이 변형한 어휘의 질의개념 연관도를 반영한 모델은 다음과 같다.

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

여기서 QConcept(t,Q)는 어휘 t 가 Q 에 포함된 어휘들 중에서 몇 개의 어휘들과 관계를 갖고 있는가를 나타내는 것이다. 값의 범위는 $0 \le QConcept(t,Q) \le |Q|$이다. |Q|는 질의 어휘 개수이다. 어휘가 질의개념과 관계를 맺고 정도는 어휘번역확률 정보를 이용하였다.

어휘 t 가 질의 Q 의 세 개의 어휘에 대해서 모두 어휘번역확률 값을 갖고 있다면 QConcept(t, Q)값은 3 이고, 두 개의 어휘에 대해서만 어휘번역확률 값을 갖고 있다면 2 가 된다. 어떤 어휘 t 가 질의 Q 에 나타난 모든 어휘와 어휘번역확률을 갖고 있다면 어휘 t 는 질의 개념을 잘 반영한다고 볼 수 있다.

(그림 2)는 TREC AP 의 학습질의에 나타나는 Q_{133} = 'hubble space telescope'에 대해서 각 어휘들이 질의와의 개념 연관을 가질 경우에 선으로 표시하였다. 어휘 'NASA'가 질의 Q_{133} = {'hubble', 'space', 'telescope'}의 세 개의 어휘에 대해서 모두 어휘번역확률 값을 가지고 있으므로 QConcept('NASA', Q_{133})의 값은 3 이 되고, 어휘 'shop'은 한 개의 어휘에 대해서만 어휘번역확률 값을 갖고 있으므로 QConcept('shop', Q_{133})는 1 이 된다.



(그림 2) 어휘의 질의개념 연관도를 반영한 예제

4. 실험 및 평가

본 논문에서 제안한 방법의 유효성을 검증하기 위하여 정보검색 실험집합인 TREC AP(88-90) 컬렉션에 대해서 언어 모델(LM), 어휘 관계를 반영한 번역기반 언어 모델(TransLM)과 어휘의 질의개념 연관도를 반영한 모델(QConceptTransLM)과 비교 실험을 하였다.

질의를 학습질의집합 100 개(Q51-Q150)와 테스트질의집합 50 개(Q151-Q200) 로 구분하여, 학습질의를 이용해서 각 모델에 필요한 파라미터를 학습하였고, 테스트질의에 대해서 평가를 하였다. 성능 평가는 MAP(Mean Average Precision)을 이용하였다.

언어모델에서의 파라미터 μ 를 설정하기 위해 다음과 같은 값 μ = {500, 1000, 2000, 2500, 3000, 5000}에서 학습질의에서 가장 좋은 성능을 보인 μ = 2000 으로 설정하였고 번역기반 언어모델에서 번역부분의 가중치를 나타내는 파라미터 β = {0.1, 0.2, …, 0.9}에서 학습질의에서 가장 좋은 값을 보인 값으로 설정했다.

실험에서는 어휘의 번역확률에서 자신으로 번역될 때(self-translation)의 확률 즉, P(w|w)의 번역값을 0 으로 둔 것, 1 로 둔 것과 번역확률 계산에서 나온 값을 그대로 사용한 것으로 구분하여 실험하였다. <표 3>과 <표 4>에서 보는 바와 같이, 어휘의 질의개념 연관도를 반영한 모델이 번역기반 언어모델에 비해 학습질의에서 우수한 성능을 보이고 있다. 이때 자기번역확률을 0 으로 두었을 때가 가장 좋다.

테스트 질의집합에 대한 성능평가 결과는 <표 5>

<표 3> 학습질의에 대한 TransLM 에서 자기번역확률에 따른 평균정확률

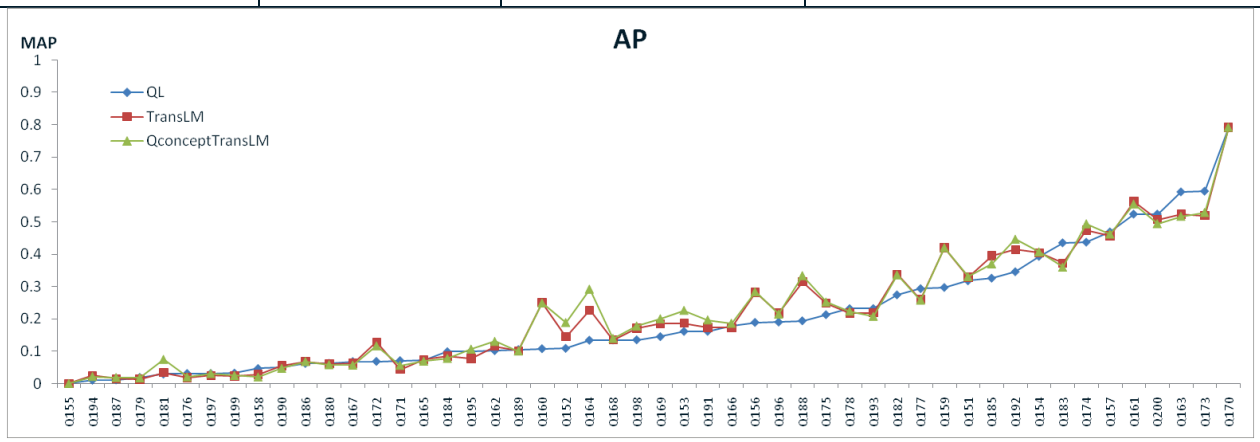
P(w|w)	LM	TransLM 에서 번역부분 β의 중요도								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.2166	0.2183	0.2198	0.2213	0.2231	0.2247	0.2268	0.2279	**0.2284**	0.2274
1		0.2180	0.2192	0.2198	0.2205	0.2211	0.2216	0.2219	**0.2221**	0.2214
번역값		0.2181	0.2196	0.2207	0.2217	0.2227	0.2238	0.2252	**0.2257**	0.2256

<표 4> 학습질의에 대한 QConceptTransLM 에서 자기번역확률에 따른 평균정확률

P(w|w)	LM	QConceptTransLM 에서 번역부분 β의 중요도								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.2166	0.2203	0.2231	0.2260	0.2283	0.2295	0.2313	**0.2315**	0.2304	0.2247
1		0.2198	0.2218	0.2236	0.2246	0.2257	0.2265	**0.2272**	0.2271	0.2260
번역값		0.2194	0.2209	0.2218	0.2224	**0.2229**	**0.2229**	0.2226	0.2220	0.2211

<표 5> TREC AP 테스트 질의집합에 대한 성능 비교

P(w|w)	LM	TransLM	QConceptTransLM
0	0.2042	**0.2193(+7.4%)**	**0.2252(+10.3%)**
1		0.2156	0.2209
번역값		0.2183	0.2173



(그림 3) 어휘의 질의개념 연관도를 반영한 예제

에 나타나 있다. 테스트 질의집합 역시 자기번역확률이 0 일 때가 가장 좋다. 번역기반 언어모델이 언어모델에 비해 7.4%의 성능향상을 보였고, 어휘의 질의개념 연관도를 반영한 모델이 언어모델에 비해 10.3%의 성능향상을 보였다.

(그림 3)은 각 질의에 대한 LM, TransLM, QConceptTransLM 의 평균정확률에 대한 성능을 보여주고 있다. 어떤 질의에 대해서는 TransLM 과 QConceptTransLM 에 의한 성능이 LM 보다 떨어지는 경향도 있음을 알 수 있다. 대부분의 질의에 대해서 QConceptTransLM 을 적용한 것이 우수한 성능향상을 보여준다.

5. 결론 및 향후연구

본 논문에서는 어휘-어휘 관계 정보와 어휘의 질의개념 연관도를 반영한 정보검색 모델을 제안하였다. 어휘-어휘 관계 정보를 획득하기 위하여 하나의 문서에서 어떤 한 문장과 그 다음 문장 사이에는 내용의 흐름에 있어서 연관성을 가지고 기술되고 있다고 가정하고, 문장-다음문장 번역쌍을 이용해서 번역확률을 계산하였다. 뉴스기사 컬렉션인 TREC AP 컬렉션에 대한 실험에서 번역기반 언어모델이 언어모델에 비해 7.4%의 성능향상을 보였고, 어휘의 질의개념 연관도를 반영한 모델이 언어모델에 비해 10.3%의 성능향상을 보였다. 이 결과는 본 논문에서 제안한 문장-문장 쌍에 대해 번역확률을 계산한 방법이 유효하고, 어휘의 질의개념 연관도를 반영하는 것이 의미 있음을 보여주었다.

향후 연구에서는 본 연구에서 제시한 {S1-S1, S1-S2} 번역쌍이 아닌 다른 여러 번역쌍에 대한 어휘번역확률을 계산하여 최적의 어휘번역확률을 추출할 수 있는 번역쌍에 대한 연구가 필요하다.

참고문헌

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

[2] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. Comput. Linguist., 19(2):263–311, 1993.

[3] V. Murdock and W. B. Croft. “A Translation Model for sentence retrieval.” In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 684–691, 2005.

[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] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In Proceedings of the 14th ACM Conference on Information nd Knowledge Management, pages 84–90, 2005.

[6] R. Jin, A. G. Hauptmann, and C. Zhai. Title language model for information retrieval. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 42–48, 2002.

[7] 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.

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