http://dx.doi.org/10.3745/KIPSTB.2012.19B.3.183

# 어휘 번역확률과 질의개념연관도를 반영한 검색 모델

김 준 길[†] · 이 경 순[††]


## 요 약

정보 검색에서 성능 저하의 주요 요인은 사용자의 질의와 검색 문서 사이에서의 어휘 불일치 때문이다. 어휘 불일치 문제를 해결하기 위해 본 논문에서는 어휘 번역확률을 이용한 번역기반 언어모델에 질의개념연관도를 반영한 검색 모델을 제안한다. 어휘관계 정보를 획득하기 위하여 문장-다음문장 쌍을 이용하여 어휘 번역확률을 계산하였다. 제안모델의 유효성을 검증하기 위해 TREC AP 컬렉션에 대해 실험하였다. 실험결과에서 제안모델이 언어모델에 비해 아주 우수한 성능향상을 보였고, 번역기반 언어모델에 비해서도 높은 성능을 나타냈다.


**키워드 : 정보검색, 어휘관계, 질의개념, 어휘 번역확률, 번역기반 언어모델**

# Retrieval Model Based on Word Translation Probabilities and the Degree of Association of Query Concept

Jun-Gil Kim[†] · Kyung-Soon Lee[††]


## ABSTRACT

One of the major challenge for retrieval performance is the word mismatch between user's queries and documents in information retrieval. To solve the word mismatch problem, we propose a retrieval model based on the degree of association of query concept and word translation probabilities in translation-based model. The word translation probabilities are calculated based on the set of a sentence and its succeeding sentence pair. To validate the proposed method, we experimented on TREC AP test collection. The experimental results show that the proposed model achieved significant improvement over the language model and outperformed translation-based language model.


**Keywords : Information Retrieval, Word Relationships, Query Concept, Word Translation Probabilities, Translation-based Language Model**

## 1. 서 론

정보검색 분야에서 어휘의 다양한 표현으로 인한 어휘 불일치 문제(Word Mismatch Problem)는 검색 성능 저하의 주요 원인이다. 이러한 어휘 불일치 문제를 해결하기 위하여 어휘관계 정보(Word Relationships)를 반영한 정보검색 모델에 관한 연구[1,2,3,4]가 활발히 이루어지고 있다. 어휘관계 정보를 반영함으로써 질의가 문서에 나타나지 않았더라도 질의와 연관된 어휘가 포함된 문서를 검색하는 것이 가능하다.

어휘관계 정보를 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 번역 모델(Translation Model)을 제안하였고, Jeon[4]은 질의응답 아카이브에서 유사한 질문을 찾는 문제에서 IBM 모델 1과 언어모델(Language Model)[5]을 결합한 번역기반 언어모델(Translation-based Language Model)을 제안하였다. 질의들 사이에 같은 어휘가 존재하지 않더라도 어휘관계 정보를 이용하여 비슷한 질의를 찾아주었다.

번역기반 언어모델에서 같은 단어 사이의 어휘번역 확률($P(t|s)$)에서의 자기번역 확률($P(t'=t|t)$)을 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[4]은 모든 어휘에 대하여 자신의 번역확률

† 준 회 원 : 전북대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 전북대학교 컴퓨터공학부 영상정보신기술연구센터 부교수(교신저자)
논문접수 : 2011년 4월 29일
수 정 일 : 1차 2011년 11월 22일, 2차 2012년 1월 25일
심사완료 : 2012년 1월 31일

을 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 Jeon[4]이 제안한 번역기반 언어모델에 번역확률 계산방법을 제안하였다. 질의와 대답을 쌍으로 묶어주어 질의와 대답을 각각 번역부분의 소스로 하였을 때의 확률 계산방법과 질의-대답쌍과 대답-질의쌍을 결합하여 확률을 계산하는 방법을 번역기반 언어모델로 비교 실험하였으며 번역기반 언어모델의 유효성을 보였다. 질의응답 아카이브에서 구축한 어휘 번역확률 정보를 정보검색에 이용하기에는 검색대상인 컬렉션의 성격이 다르다는 문제점이 있다. 또한, 연구[8]은 질문 분류에서의 어휘 불일치 문제를 해결하기 위해 같은 범주에 속하는 질문-질문 쌍들에 대해서 번역확률을 계산하여 번역기반 언어모델로 질문을 분류하는 방법을 제안하였다.

본 논문에서는 일반적인 정보검색에서 (i) 검색대상인 문서에 대해서 번역확률을 계산하는 방법을 제안한다. 번역확률 계산에서 한 문장과 이후에 인접하게 나타나는 문장들 사이에는 내용의 흐름에 있어서 연관성이 있다는 것을 가정하고, 문장 사이의 어휘 번역확률을 계산하였다. 또한, (ii) 기존의 어휘관계 정보를 반영하는 번역기반 언어모델에 추가적으로 어휘와 질의 개념과의 연관 정도를 반영한 모델을 제안한다. 어휘의 질의개념(Query Concept) 연관도는 어휘가 질의가 내포하는 개념과 어느정도 연관성이 있는지를 반영하는 것이다. TREC AP 테스트 컬렉션을 이용하여 제안 방법의 유효성을 검증한다.

본 논문의 구성은 2장에서 어휘 번역확률 정보 획득 방법에 대하여 설명하고, 3장에서는 번역기반 언어모델과 어휘의 질의개념연관도를 반영한 번역기반 언어모델에 대하여 설명하고, 4장에서는 실험결과를 보여주고, 5장에서는 결론을 내린다.

## 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 타겟(target)으로 보고 번역확률을 계산하게 된다. P(t|s)는 소스 어휘가 s일 때 타겟 어휘가 t일 번역 확률을 나타낸다. 어휘-어휘 번역확률은 두 어휘 사이에 관계가 있음을 나타내고, 그 관계의 중요도를 번역확률 값으로 나타낸다. 본 논문에서는 소스와 타겟이 서로 같은 언어쌍에 대하여 번역확률을 계산하였다.

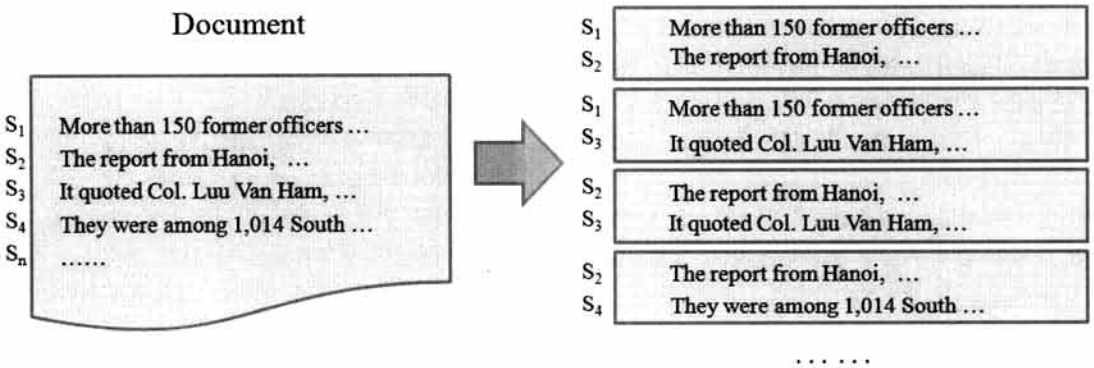
어휘-어휘 사이의 관계정보를 얻기 위해서 본 논문에서는 하나의 문서 내에서 어떤 한 문장과 그 이후에 인접하게 나타나는 문장들 사이에는 내용의 흐름에 있어서 연관성을 가지고 기술되어 있다고 가정한다. 문서에서 나타난 문장을 {S1, S2, …, Sn}이라 할 때, 문장 S1과 그 다음 문장 S2, 그 다음 문장인 S3에 나타난 어휘들도 서로 연관이 있다고 보고 {S1-S2, S1-S3} 번역쌍으로 표현해서 전체 컬렉션에 있는 번역쌍들을 이용해서 학습을 한다. 이때 S1은 소스 문장이고 S2와 S3는 타겟 문장이다. 검색대상 문서집합인 TREC AP 컬렉션에서 번역확률을 계산하였다.

문서에서 이웃 문장쌍을 이용한 번역쌍 표현은 (그림 1)과 같이, 자신의 문장과 다음 문장의 번역쌍과 자신의 문장 이후 두 번째에 나타난 문장을 번역쌍으로 표현한 {S1-S2, S1-S3} 번역쌍으로 실험하였다. 어휘 번역확률 계산을 위한 TREC AP 컬렉션에서의 문장 번역쌍 개수는 <표 1>과 같다.

〈표 1〉 어휘 번역확률 계산을 위한 문장 번역쌍 정보

| 문서 개수 | 문장 개수 | 유일한 단어 개수 | 문장 번역쌍 개수 |
|---|---|---|---|
| 242,918 | 5,231,527 | 246,307 | 9,622,036 |

어휘-어휘 번역확률을 계산하는 간단한 방법은 두 어휘의 공기빈도(co-occurrence)를 이용하는 것이다. 만일 한 타겟 어휘 t가 주어진 소스 어휘 s와 여러 번 같이 나타났다면, 이 타겟 어휘 t는 소스 어휘 s의 번역 어휘가 될 확률이 아주 높다. 그러나 이 방법을 단순히 적용하면 대량의 불용어에 큰 확률을 주게 된다. 그것은 불용어가 소스 어휘 s와




(그림 1) 문서에서 이웃 문장 번역쌍을 표현한 예제

〈표 2〉 이웃 문장 번역쌍을 이용한 어휘 번역확률에서의 일부 예제

| | 번역확률에서 타겟 부분 어휘 | | | | | |
|---|---|---|---|---|---|---|
| | hubble | | space | | telescope | |
| 소스 부분 어휘 | spaceborne | 0.1407 | cosmonaut | 0.1934 | telescope | 0.1656 |
| | weiler | 0.1136 | space | 0.1770 | weiler | 0.1474 |
| | hubble | 0.1001 | mir | 0.1661 | spectrograph | 0.1075 |
| | westphal | 0.0998 | nasa | 0.1324 | nebula | 0.0643 |
| | telescope | 0.0567 | apollo | 0.0846 | nasa | 0.0085 |
| | starlight | 0.0180 | shuttle | 0.0570 | shuttle | 0.0033 |
| | shuttle | 0.0097 | color | 0.0005 | shop | 0.0007 |
| | nasa | 0.0052 | worldwide | 0.0002 | worldwide | 0.0001 |

자주 함께 나타나기 때문이다. 따라서 의미 없는 공동 발생을 배제하는 더욱 좋은 방법이 필요하다.

본 논문에서는 어휘 번역확률을 EM 알고리즘을 이용한 GIZA++[9, 10]를 이용하여 번역 확률을 계산하였다. GIZA++은 통계적 번역 모델 툴킷으로서 여러 번역 모델을 지원해준다. 문장-문장 번역쌍을 이용한 어휘 번역확률 계산 예제는 <표 2>와 같다. 학습질의 'hubble space telescope'에 나타난 어휘에 대한 번역확률을 나타냈다.

## 3. 어휘의 질의개념연관도를 반영한 번역기반 언어모델

### 3.1 번역기반 언어모델

Jeon[4]과 Xue[7]는 어휘관계 정보를 정보검색에 반영하기 위하여 언어모델(Language Model : LM)과 어휘 번역확률을 반영하는 IBM 모델 1[2]을 개선한 번역기반 언어모델(Translation-based Language Model : TransLM)[4,6]을 사용하였다.

번역기반 언어모델은 언어모델에 수식 (3)에서와 같이 문서에 나타난 어휘 t와 질의에 나타난 어휘 q의 번역확률을 의미하는 P(q|t)를 반영한 것이다.

$$P(Q|D)=\prod_{q\in Q}P(q|D) \tag{1}$$

$$P(q|D)=\frac{|D|}{|D|+\mu}P_{mx}(q|D)+\frac{\mu}{|D|+\mu}P_{ml}(q|C) \tag{2}$$

$$P_{mx}(q|D)=(1-\beta)P_{ml}(q|D)+\beta\sum_{t\in D}P(q|t)P_{ml}(t|D) \tag{3}$$

$$P_{ml}(q|D)=\frac{freq(q,D)}{|D|},P_{ml}(q|C)=\frac{freq(q,C)}{|C|} \tag{4}$$

여기서 Q는 질의를 나타내고 D는 문서, C는 컬렉션을 의미한다. |D|는 문서에 나타난 어휘의 개수이고 |C|는 컬렉션에 나타난 어휘의 총 개수가 되고, freq(q, D)는 어휘 q가 문서 D에 나타난 빈도수이고 freq(q, C)는 어휘 q가 컬렉션 C에 나타난 빈도수이다. 수식 (3)에서의 $P_{mx}$은 번역기반 언어모델에서의 수식이고, 수식 (4)에서의 $P_{ml}$은 언어모델에서의 수식이다. 수식 (2)에서 $P_{mx}(q|D)$가 $P_{ml}(q|D)$가 된다면 언어모델이 된다.

번역기반 언어모델에서 어휘 번역확률 정보를 반영시키는 수식 (3)에서 P(q|t)는 문서에 나타난 어휘 t가 질의 어휘 q로 번역되는 확률이다. 질의에 나타난 어휘 q가 문서에 나타나지 않더라도 문서에 나타난 어휘 t와 q와의 관계를 이용하여 검색에 반영한다. β값으로 어휘 번역확률 부분의 중요도를 조정할 수 있다. 만약 β에 0을 부여한다면 언어모델이 된다.

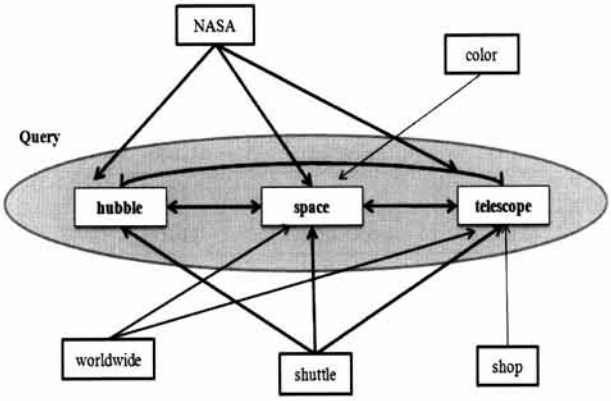
### 3.2 질의개념연관도를 반영한 번역기반 언어모델

본 논문에서는 번역기반 언어모델에 어휘와 질의 개념과의 연관 정도를 반영한 모델(QConceptTransLM)을 제안한다. 위 수식(3)을 다음과 같이 변형한 어휘의 질의개념연관도를 반영한 모델은 다음과 같다.

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

여기서 QConcept(t,Q)는 어휘 t가 Q에 포함된 어휘들 중에서 몇 개의 어휘들과 관계를 갖고 있는가를 나타내는 것이다. 값의 범위는 0≤QConcept(t,Q)≤|Q|이다. |Q|는 질의 어휘 개수이다. 어휘가 질의개념과 관계를 가지는 정도는 어휘 번역확률 정보를 이용하였다.

어휘 t가 질의 Q의 세 개의 어휘에 대해서 모두 어휘 번역확률 값을 갖고 있다면 QConcept(t, Q)값은 3이 되고, 두 개의 어휘에 대해서만 어휘 번역확률 값을 갖고 있다면 2가 된다. 어떤 어휘 t가 질의 Q에 나타난 모든 어휘와 어휘 번역확률을 갖고 있다면 어휘 t는 질의 개념을 잘 반영한다고 볼 수 있다.




(그림 2) 어휘의 질의개념연관도를 반영한 예제

(그림 2)는 TREC AP 의 학습질의에 나타나는 Q133 = 'hubble space telescope' 에 대해서 각 어휘들이 질의와의 개념 연관을 가질 경우에 선으로 표시하였다. 어휘 'NASA'가 질의 Q133 = {'hubble', 'space', 'telescope'} 의 세 개의 어휘에 대해서 모두 어휘 번역확률 값을 가지고 있으므로 QConcept('NASA', Q133)의 값은 3이 되고, 어휘 'shop'은 한 개의 어휘에 대해서만 어휘 번역확률 값을 갖고 있으므로 QConcept('shop', Q133)는 1이 된다. 따라서, 질의 Q133 = 'hubble space telescope' 과 의미적으로 연관성이 높게 나타나는 'NASA', 'shuttle'과 같은 어휘에 대해 높은 가중치를 부여하게 된다. 결과적으로 이전의 어휘연관성을 고려한 방법인 TransLM에서는 단순히 각각의 질의 어휘들과 연관성이 있는 어휘들에 대한 가중치만 부여하였지만, 제안한 QConceptTransLM에서는 추가적으로 질의 전체의 의미를 내포하고 있는 어휘에 대해 좀 더 높은 가중치를 부여하여 전체 질의 의미를 검색에 반영하게 된다.

## 4. 실험 및 평가

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

- LM : 기존의 검색 모델
- TransLM : 기존의 어휘번역확률을 이용하는 검색 모델
- QConceptTransLM : TransLM에 추가적으로 제안한 질의와의 연관도를 반영한 검색 모델

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

$$MAP = \frac{1}{|Q|}\sum_{q \in Q} ap(q) \qquad (6)$$

여기서 ap(q)는 질의 집합 Q에 있는 질의 q에 대한 평균 정확률을 나타낸다.

언어모델에서 수식 (2)의 파라미터 μ는 {500, 1000, 2000, 2500, 3000, 5000}에서 학습질의에서 가장 좋은 성능을 보인 2000으로 설정하였다. 번역기반 언어모델에서 번역부분의 가중치를 나타내는 수식 (3)의 파라미터 β는 {0.1, 0.2, …, 0.9}에서 학습질의에서 가장 좋은 성능을 보인 값으로 설정했다.

번역확률 계산방법에 따라서 성능에 변화가 생기므로 본 실험에서는 어휘의 번역확률에서 자신으로 번역될 때(self-translation)의 확률 즉, P(t'=t|t)의 번역값을 0으로 둔 것, 1로 둔 것과 번역확률 계산에서 나온 값을 그대로 사용한 것에 대해 비교실험하였다.

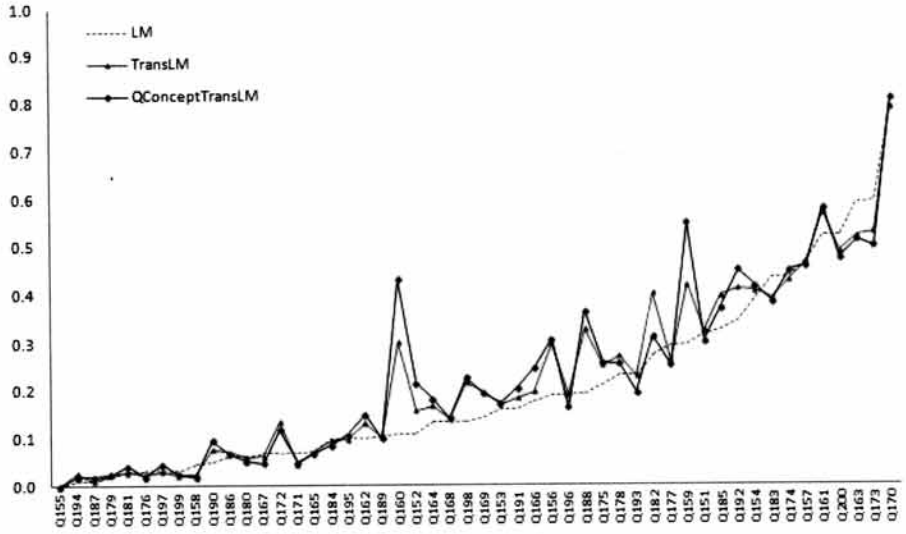
〈표 3〉 TREC AP 테스트 질의집합에 대한 성능 비교

| P(t'=t\|t) | LM | TransLM | QConceptTransLM |
|---|---|---|---|
| 0 | 0.2042 | 0.2218 (+8.6%) | 0.2278 (+11.5%) |
| 1 | | 0.2171 (+6.3%) | 0.2144 (+4.9%) |
| 번역값 | | 0.2257 (+10.5%) | **0.2314 (+13.3%)** |

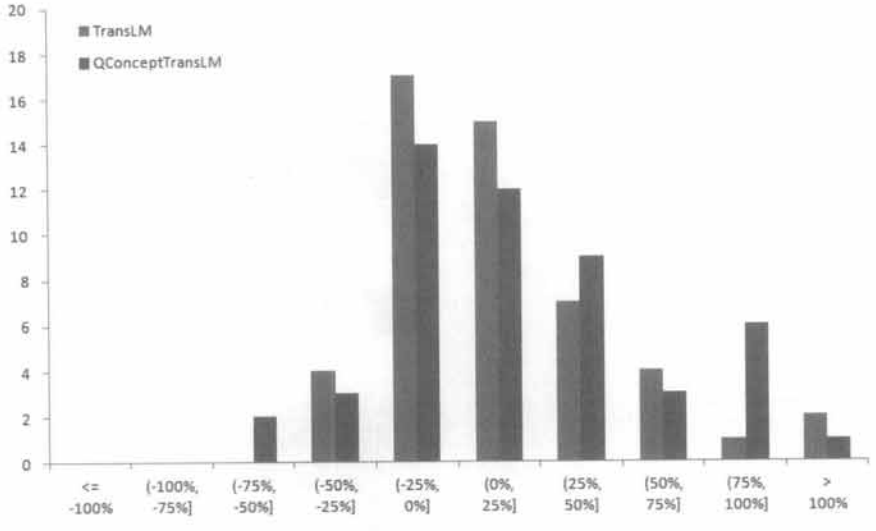
테스트 질의집합에 대한 성능평가 결과는 <표 3>에 나타나 있다. 성능향상은 언어모델(a)을 기준으로해서 (b - a) / a × 100으로 계산하였다. 자기 번역확률을 번역값으로 하였을 때 성능이 가장 좋았고, 언어모델에 비해 번역기반 언어모델이 10.5%의 성능향상을 보였으며, 어휘의 질의개념연관도를 반영한 모델이 13.3%의 성능향상을 보였다.

(그림 3)은 두 어휘 번역확률에서의 각 질의에 대한 LM, TransLM과 QConceptTransLM의 평균정확률에 대한 성능을 보여주고 있다. 어떤 질의에 대해서는 TransLM과 QConceptTransLM에 의한 성능이 LM보다 떨어지는 경향도 보였다. 전체 50개의 질의에 대해서 어휘 번역확률을 이용한 TransLM이 LM에 비해 29개의 질의에서 성능향상을 보였고 21개의 질의에서는 성능이 떨어짐을 알 수 있다. 또한 질의개념연관도를 반영한 QConceptTransLM이 LM에 비해 31개의 질의에서 성능향상을 보였고 19개의 질의에서 성능이 떨어짐을 알 수 있다. TransLM에서 성능향상을 보인 29개의 질의에서 QConceptTransLM이 20개의 질의에서 더 높은 성능향상을 보였다.

(그림 4)에서는 TransLM과 QConceptTransLM의 성능분포를 나타낸다. TransLM에 비해서 QConceptTransLM이 성능향상을 보인 질의의 수가 더 많은 것을 알 수 있다. (그림 3)과 (그림 4)에서 어떤 질의에 대해서 낮은 성능을 보이는 경향이 있는데 질의 확장에서 질의를 확장하였을 경우 성능이 떨어지는 경우와 유사하게 주변 단어에 가중치를 부여하지 않을 경우에 더 나은 성능을 보이는 경우가 존재한다는 것을 알 수 있다. 이 부분에 대해서는 향후 연구 과제로 남겨놓도록 하겠다.




(그림 3) 각 질의에 대한 LM, TransLM과 QConceptTransLM의 평균정확률에 대한 성능 비교




(그림 4) 성능분포에서 TransLM과 QConceptTransLM의 질의 개수에 대한 비교

〈표 4〉 높은 성능향상을 보인 질의에서 상위 10개 문서에서의(P@10) 비교

| Query | MAP | | | P@10 | | |
|---|---|---|---|---|---|---|
| | LM | TransLM | QConceptTransLM | LM | TransLM | QConceptTransLM |
| Q152 | 0.109 | 0.158 | 0.215 | 0.700 | 0.600 | 0.900 |
| Q160 | 0.108 | 0.301 | 0.432 | 0.200 | 0.500 | 0.700 |
| Q192 | 0.345 | 0.412 | 0.451 | 0.400 | 0.600 | 0.900 |

<표 4>에서는 TREC AP 테스트 질의집합에서 높은 성능향상을 보인 질의의 일부에 대해서 상위에 검색된 문서의 정확률을 보여준다. P@n은 상위 n개의 문서에서의 정확률을 나타낸다. 예를들어, P@10은 상위 10개의 문서에서 정답 문서 3개가 포함되었을 경우에는 값이 0.3이 된다. 상위 10개의 문서에 존재하지 않았던 정답 문서가 상위 10개의 문서에 추가적으로 포함됨으로써 LM을 사용했을 때보다 TransLM과 QConceptTransLM을 사용했을 때에 각각의 질의에서의 P@10 성능이 향상됨을 알 수 있다.

## 5. 결 론

본 논문에서는 일반적인 정보검색에서 검색대상인 문서에 대해서 번역확률을 계산하는 방법을 제안하였다. 어휘-어휘 관계 정보를 획득하기 위하여 하나의 문서에서 어떤 한 문장과 그 이후에 나타난 문장 사이에는 내용의 흐름에 있어서 연관성을 가지고 기술되고 있다고 가정하고, 문장-두 개의 다음 문장 번역쌍의 이웃 문장 번역쌍을 이용해서 번역확률을 계산하였다.

또한, 기존의 어휘관계 정보를 반영하는 번역기반 언어모델에 추가적으로 어휘와 질의 개념과의 연관 정도를 반영한 모델을 제안하였다. 이러한 어휘 번역확률을 이용함으로써 문서에 질의 어휘가 나타나지 않았더라도 주변에 나타난 질의와 연관된 어휘를 통해 문서에 대한 가중치를 부여해 주었고, 추가적으로, 제안한 어휘의 질의개념연관도를 반영한 정보검색 모델을 이용하여 질의와 연관성이 높은 어휘들에 대해 좀 더 높은 가중치를 부여하여 검색에 반영하였다.

뉴스기사 컬렉션인 TREC AP 컬렉션에 대한 실험에서 언어모델에 비해 번역기반 언어모델이 10.5%의 성능향상을 보였고, 어휘의 질의개념연관도를 반영한 모델이 13.3%의 성능향상을 보였다. 이 결과는 본 논문에서 제안한 이웃 문장쌍에 대해 번역확률을 계산한 방법이 유효하고, 어휘의 질의개념연관도를 반영하는 것이 유효함을 보여주었다.

향후 연구에서는 관계가 있는 어휘들 사이에는 높은 번역확률을 가지면서 관계가 없는 어휘들 사이에는 번역 확률을 가지지 않게 하는 안정된 어휘 관계에 대한 번역 확률을 구축하는 방법에 대한 연구가 필요하다.

## 참 고 문 헌


[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," Computational Linguistics 19(2), pp.263-311, 1993.

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

[4] J. Jeon, W. B. Croft and J. H. Lee, "Finding Similar Questions in Large Question and Answer Archives," Proceedings of the 14th ACM CIKM Conference, pp.84-90, 2005.

[5] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.275-281, 1998.

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

[7] X. Xue, J. Jeon and W. B. Croft, "Retrieval Models for Question and Answer Archives," Proceedings of the 31st annual international ACM SIGIR conference, pp.475-482, 2008.

[8] 김설영, 이경순, "질문대답 아카이브에서 어휘 연관성을 이용한 질문 분류," 정보처리학회논문지B, 제17권 제4호, pp.327-332, 2010.

[9] GIZA tool. http://code.google.com/p/giza-pp/

[10] F. J and Och, H. Ney. "A Systematic Comparison of Various Statistical Alignment Models," Proceedings of the Computational Linguistics, Vol.29, No.1, pp.19-51, 2003.



**김 준 길**

e-mail : jgkim@jbnu.ac.kr

2011년 전북대학교 컴퓨터공학과(학사)

2011년~현 재 전북대학교 컴퓨터공학과 석사과정

관심분야 : 정보검색, 정보 마이닝, 자연언어처리



**이 경 순**

e-mail : selfsolee@jbnu.ac.kr

1994년 계명대학교 컴퓨터공학과(학사)

1997년 한국과학기술원 전자전산학(석사)

2001년 한국과학기술원 전자전산학(박사)

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

2007년 미국 매사추세츠주립대학 방문교수

2004년~현 재 전북대학교 컴퓨터공학부 영상정보신기술 연구센터 부교수

관심분야 : 정보검색, 정보마이닝, 자연언어처리