한-일 교차언어검색에서의 질의 문맥 정보를 이용한 대역어 변환 확률 모델

이규찬 [0], 강인수, 나승훈, 이종혁
포항공과대학교 지식 및 언어공학 연구실
{117690[0], dbaisk, nsh1979, jhlee} @postech.ac.kr

Query Context Information-Based Translation Models for Korean-Japanese Cross-Language Information Retrieval

Gyu-Chan Lee[0], In-Su Kang, Seung-Hoon Na, Jong-Hyeok Lee
Knowledge & Language Engineering Lab, POSTECH

요 약

교차언어 검색 과정에서는 질의나 문서의 언어를 일치시키기 위한 변환 과정이 필수적이며, 이런 변환 과정에서 어휘의 중의성으로 인해 하나의 어휘에 대응하는 다수의 대역어가 생성됨으로써 사용자의 정보 욕구를 왜곡시켜 검색의 성능을 저하시킬 수 있다. 본 논문에서는 어휘 중의성 문제를 해결하기 위해서 질의의 문맥 정보를 이용하여 변환 질의의 확률을 구함으로써 중의성을 해소하는 방식을 제시하고, 질의의 길이, 중의도, 중의성을 가진 어휘의 비율 등에 따라서 성능이 어떻게 변하는지 비교함으로써 이 방법의 장점과 단점을 분석한다. 또한 현재의 단점을 보완하기 위한 차후 연구 방향을 제시한다.

1. 서론

교차언어검색은 질의와 문서집합의 언어가 서로 다른 상황에서의 검색을 의미한다[9]. 이러한 상황에서 검색을 가능하게 하기 위해서는 질의와 문서집합의 언어를 서로 일치시키는 변환 과정이 필요하다.

서로 다른 언어를 일치시키는 변환 과정을 위해서는 반드시 언어 자원을 필요로 하게 된다. 주로 사용되는 언어 자원으로는 기계 번역기, 병렬 코퍼스, 전자 사전 등이 있다. 기계 번역기는 구축이 힘들고 번역에 많은 자원을 소모할뿐더러 유사한 언어 사이가 아닐 경우 효용성이 높지 않다는 문제점을 지니고 있다. 병렬 코퍼스는 자원이 충분히 주어질 경우 단일어 검색에 버금가는 성능을 낼 수 있지만[12] 충분한 양의 코퍼스를 얻기 힘들다는 점과 병렬 코퍼스 간의 문단, 문장, 어휘 간의 정렬이 힘들다는 단점을 가지고 있다.

이런 이유로 인해서 가장 흔하게 사용되는 자원은 전자 사전이다. 전자 사전은 상대적으로 구축이 쉽고 적용이 간단하다는 장점이 있지만, 변환 과정에서 어휘 중의성 문제를 일으키게 되므로, 중의성 처리가 중요한 이슈가 된다[1, 2, 3, 4, 8, 10, 11, 13].

어휘 중의성이란 자연언어의 어휘들이 여러 개의 의미를 가지는 현상을 가리킨다. 사전을 이용해서 언어를 변환할 경우, 중의성을 가진 어휘가 나타난다면 사전은 이 어휘에 대해 다수의 대역어를 생성하게 되고, 이렇게 생성된 대역어 집합에는 올바른 대역어와 그렇지 않은 어휘들이 혼재하게 된다. 질의나 문서의 원래 의미와는 상관 없는 어휘는 내용을 왜곡시키는 잡음 효과를 발생시켜서 검색의 성능을 떨어뜨리는 중요한 요인으로 작용하게 된다.

언어 변환의 대상은 질의 혹은 문서 모두 가능한데 주로 길이가 짧아 변환 시간이 오래 걸리지 않는 질의 변환 방식이 선호되고 있다. 그러나 질의에서는 중의성을 처리하기에 충분한 문맥정보를 얻기 힘들다는 단점이 있기 때문에, 상대적으로 적은 문맥정보를 이용해서 효율적으로 중의성을 처리할 수 있는 방법의 연구가 중요하다.

이 논문에서는 문맥 정보를 이용한 대역어의 변환 확률 모델을 제시한다. 그리고 이 모델을 이용한 실험 결과를 다른 방법들과 질의의 길이와 질의의 중의도, 중의성을 가진 어휘의 비율 등에 따라서 비교 분석한다. 그리고 이를 토대로 드러난 문제점을 해결하기 위한 방법을 제시할 것이다.

2. 관련 연구

언어를 변환하는 과정에서 발생하는 중의성은 교차언어 검색의 성능이 단일언어 검색의 성능보다 떨어지도록 하는 중요한 요인이다. 따라서 대역어의 중의성을 해소하기 위한 다양한 연구가 진행되었다.

핀란드의 Pirkola는 98년, 핀란드어-영어 교차언어 검색에서 하나의 질의 어휘로부터 얻은 대역어들을 그룹화하는 방식을 제안했다[10]. 하나의 그룹을 하나의 어휘처럼 사용해서 검색을 수행함으로써 30%~50% 정도의 성능 향상을 나타내었다고 보고했다. [11]에서는 이 방식을 다른 언어 쌍으로 확장, 역시 16%~43%의 성능 향상을 보였다. 이 방법은 [2]에 의해 다양한 TF와 IDF 공식이 적용되게 되었다.

다른 방향에서는 공기 정보를 바탕으로 중의성을 해결하려는 다양한 연구가 진행되었다[1, 3, 4, 13]. 이 방식은 문서 집합에서 많이 공기하는 대역어들이 정확한 대역어일 확률이 높다는 가정에 기반하고 있다. 각 연구마다 다양한 수식의 공기 정보를 이용해서 중의성을 해소한다. 특히 [3]에서는 바로 뒤에 나타나는 어휘와 의존관계를 설정, 의존 관계에 있는 어휘 사이의 공기 정보만을 계산해서 복잡도를 줄이면서 동시에 보다 정확한 중의성 해소가 가능하도록 했다.

그 외에 병렬 코퍼스를 이용해서 중의성을 해소하는 연구[7]와 어휘 대신 구 단위로 중의성을 해소하는 연구[6] 등이 있다.

[14]에서는 하나의 질의 어휘에 대응하는 대역어의 개수에 반비례하는 가중치를 부여하는 방식을 제안했다. 이 방식은 다수의 대역어를 가지는 어휘가 생성한 대역어에 낮은 점수를 부여함으로써 올바르지 않은 대역어로 인해 생기는 잡음 효과를 최소화하고 상대적으로 대역어 개수가 적은 어휘가 강조되도록 하는 효과를 가진다. 이 방법은 공기 정보를 이용해서 중의성을 해소한 경우와 큰 차이가 없는 성능을 나타냈다.

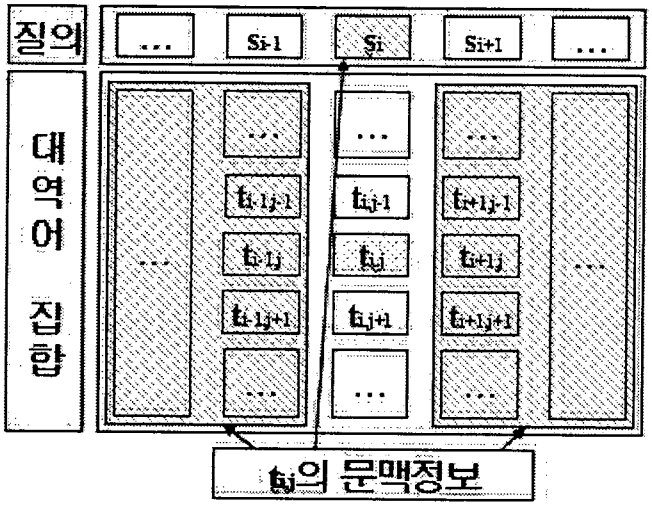


Figure 1: 문맥 정보

3. 문맥 정보를 이용한 대역어 변환 확률 모델

대역어 확률 모델은 질의의 문맥 정보에 따라서 현재 대역어가 올바른 대역어일 확률을 알 수 있다는 가정에서 출발한다. 문맥정보는 현재 대역어의 근원이 되는 질의 어휘와 주변 대역어들을 포함한다.

주어진 질의의 i번째 어휘에 대한 여러 대역어들 중 j번째 대역어를 t_{ij}라고 했을 때 t_{ij}가 올바른 대역어일 확률은 초기 가정에 의해서 다음과 같이 표현할 수 있다.

$$P(t_{ij} \mid s_i, T) \tag{1}$$

여기서 T는 s_i로부터 얻어진 대역어들을 제외한 나머지 모든 대역어들의 집합 나타낸다.

$$P(t_{ij} \mid s_i, T) = \frac{P(t_{ij}) \cdot P(s_i, T \mid t_{ij})}{P(s_i, T)} \tag{2}$$

(1)의 식을 베이지안 규칙에 따라 전개할 경우 (2)의 식을 얻을 수 있다.

(2)의 분모를 이루고 있는 $P(s_i, T)$는 정규화하는 역할을 수행할 뿐, t_{ij}의 확률에는 영향을 미치지 않으므로 생략할 수 있다. s_i와 T가 서로 독립이라고 가정한다면 (3)의 수식을 얻을 수 있다.

$$P(t_{ij}) \cdot P(s_i \mid t_{ij}) \cdot \prod_{t_k \in T} P(t_k \mid t_{ij}) \tag{3}$$

(3)의 수식에 log를 취할 경우 다음과 같은 수식을 얻을 수 있다.

$$\log P(t_{ij}) + \log P(s_i \mid t_{ij}) + \sum_{t_k \in T} \log P(t_k \mid t_{ij})$$

$$= \log P(t_{ij}) + \log P(s_i \mid t_{ij}) + \frac{\sum_{t_k \in T} \log P(t_k, t_{ij})}{P(t_k)} \tag{4}$$

(4)의 수식에 의하면 문서 집합 전체에서 현재 대역어가 얼마나 많이 등장하는지, 그 대역어가 다시 원래 질의의 언어로 변환될 때 원래 어휘로 다시 번역될 확률이 얼마인지, 그리고 한 문서에서 현재 대역어가 주변 문맥에 나타나는 대역어들과 얼마나 자주 공기하는지의 세 가지 요소가 영향을 미친다고 볼 수 있다. 특히 다른 두 요소에 비해서 공기 정보가 미치는 영향이 가장 크다는 점은, 기존의 공기 정보를 이용한 중의성 해소의 기본 아이디어와 일맥상통하는 결과라고 볼 수 있다.

두번째 요소는 질의의 언어로 된 문서의 색인이 있을 경우 그 색인에 나오는 정보를 이용하는 것이 정확하겠지만, 질의 언어로 된 문서 색인이 없을 경우에는 질의 변환에 사용한 사전의 표제어와 대역어를 반전시켜서 만든 사전을 이용해서 간단히 구할 수도 있다.

4. 실험 환경

실험을 위한 문서 집합으로는 NTCIR-4 일본어 문서를 사용하였다. 문서 집합은 1998년부터 1999년 2년간의 일본의 마이니치 신문의 기사 220078건과 요미우리 신문의 기사 373558건, 도합 593636건의 문서로 이루어져 있다. 문서 색인은 어휘를 기반으로 만들어졌다.

질의는 NTCIR-4에서 제공된 질의 60개 중에서 일본어 문서들 중에 해당 내용이 없는 질의 5개를 제외한 55개의 질를 사용했다. 질의는 한, 중, 일, 영의 4가지 언어로 구성되어 있고, 각 질의는 전문가에 의해서 동일한 내용으로 만들어졌다. 질의는 Title, Description, Narrative, Concept의 4개의 필드로 이루어져 있으며, 이 중 짧은 질의로는 Description, 긴 질의로는 Narrative을 사용했다.

질의 변환을 위해서 사용한 한-일 사전은 포항공대 지식 및 언어공학 연구실에서 개발한 Cobalt 한-일 번역기에 사용되는 한-일 사전을 이용하였다. 사전의 크기가 큰 것과 작은 것 두 가지가 있었으나 사전의 크기가 너무 클 경우 오히려 검색 성능을 떨어뜨리는 효과를 가져왔기 때문에 사전의 크기가 작은 쪽을 사용했다.

문서 순위 함수로는 Okapi BM25 모델을 사용했으며, 피드백을 포함한 성능을 더 끌어올릴 수 있는 다른 기법들은 사용하지 않았다.

5. 실험 과정

실험은 다음과 같은 과정을 거쳐서 이루어졌다.

1. 한국어 질의에서 어휘를 추출해낸다. 한국어는 어절과 어휘가 일치하지 않으므로 어휘를 추출해내기 위해서는 형태소 분석기를 사용하거나 각 어절마다 모든 길이의 부분문자열(substring)을 추출해내는 방법을 사용해야 한다. 부분문자열 추출 방식은 '대학생' 이라는 단어가 있다면 {대, 학, 생, 대학, 학생, 대학생}의 6개의 모든 길이의 부분 문자열을 추출해내는 방식이다. 이 방식은 많은 부정확한 어휘를 만들어내는 문제점이 있지만 합성명사를 분해하는 효과가 있고 체언을 모두 인식할 수 있다는 장점으로 인해서 형태소 분석기를 사용하는 방식에 비해서 검색 성능이 우수한 면을 보인다.

2. 얻어진 부분문자열들을 한-일 전자사전에 통과시킨다. 만약 현재 부분문자열이 사전에 있다면 대응하는 일본어 어휘들이 하나의 그룹으로 변환 질의에 포함된다. 사전에 존재하지 않는 부분문자열은 어휘가 아닌 것으로 간주한다.

3. 각 그룹에 속한 대역어의 변환 확률을 위의 모델을 통하여 계산한 후, 하나의 그룹에서 나온 대역어들의 변환 확률의 합이 1이 되도록 정규화한다.

6. 실험 및 결과 분석

위의 과정을 거쳐 만들어진 변환 질의를 통해서 NTCIR-4 환경에서 제안된 모델을 실험해 보았다.

비교 대상으로 제시된 실험의 내용은 다음과 같다.

(1) **J-J**: 일본어 단일어 검색이다. 일본어 질의에서 어휘를 추출해서 검색 질의로 사용했다. 그 외의 검색 환경은 다른 방법들과 동일하다. 교차언어검색 성능은 단일언어검색 성능을 뛰어넘기 힘들다는 점에서 일종의 상한선이라고 볼 수 있다.

(2) **KJ-Naive**: 과정 2에서 얻어진 변환 질의에 추가적인 가중치 계산 없이 그대로 사용한 가장 단순한 방식이고, 이 실험의 베이스라인이라고 볼 수 있다.

(3) **KJ-Nor**: [14]에서 제시된 방식을 통해서 대역어에 가중치를 부여하는 방식이다. 중의도가 낮은 어휘로부터 나온 대역어일수록 높은 가중치를, 중의도가 높은 어휘로부터 나온 대역어일수록 낮은 가중치를 받도록 하는 방식이다.

(4) **KJ-WSD**: 위에서 제시된 모델을 통해서 변환 확률을 구하고 가중치를 부여한 경우이다.

6.1 질의 특성 분석

변환 확률 계산의 정확성은 질의가 가지는 몇 가지 특성에 영향을 받을 수 있다. 충분한 문맥 정보 제공의 여부가 될 수 있는 질의의 길이, 공기 정보의 유효성의 기준이 될 수 있는 평균 중의도와 중의성을 가진 어휘의 비율 등에 따라 성능이 영향을 받을 수 있다. 따라서 실험에 들어가기에 앞서 질의의 이런 세 가지 특성을 조사한 결과는 다음과 같다.

	평균 길이	평균 중의도	중의어휘 비율
T	6.07	1.2066	0.1826
D	11.95	1.2679	0.2100
N	89.15	1.2476	0.1931
C	20.58	1.2474	0.2032
TDNC	127.75	1.2475	0.1958

표1: 질의별 특성 비교

6.2 전체 성능 평가

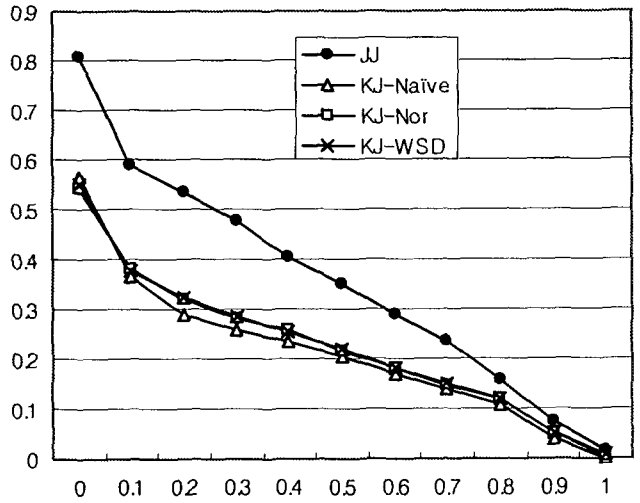
각 검색 방식을 통해 얻어진 평균 정확률은 다음과 같다.

	JJ	KJ-Naive	KJ-Nor	**KJ-WSD**
T	0.3277	0.2038	0.2080	**0.2096**
D	0.3360	0.1935	0.2104	**0.2116**
N	0.4031	0.2060	0.2707	**0.2732**
C	0.3161	0.2057	0.2126	**0.2139**
TDNC	0.4279	0.2216	0.2858	**0.2890**

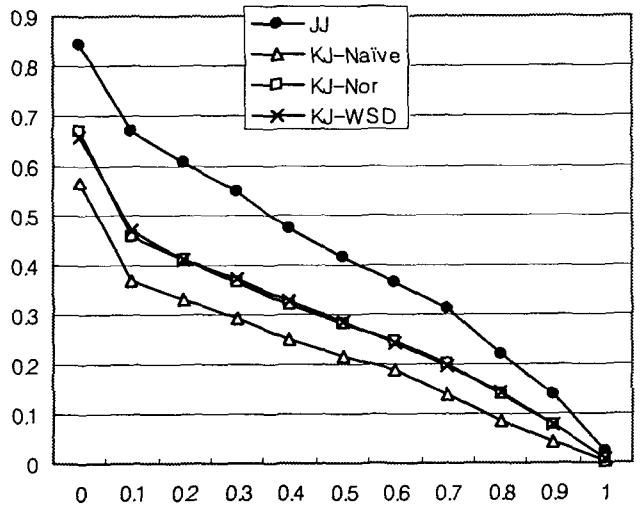
표2: 각 방식에 따른 평균 정확률

Description과 Narrative에 대한 11개 재현율 지점에서의 평균 정확률은 아래 그래프와 같이 나타낼 수 있다.

그래프1: Description



그래프2: Narrative



KJ-Nor 방식과 KJ-WSD 방식 모두 KJ-Naïve 방식에 비해 높은 성능을 보였다. 특히 질의가 길고 중의성이 높은 Narrative 필드의 경우는 32% 정도 성능이 올라가는 것을 확인할 수 있었다. 그러나 상대적으로 길이가 짧은 Title, Description, Concept 필드에서는 성능 향상 폭이 미미했다. 이를 통해서 정확한 대역어를 찾아내기 위해서는 충분히 긴 질의가 필요하며, 상대적으로 문맥 정보가 부족한 짧은 질의를 위해서 복잡한 계산을 수행해도 성능에 큰 영향을 주지 못한다는 결론을 내릴 수 있다.

다른 중요한 점은 KJ-Nor과 KJ-WSD 사이에 이렇다 할 성능 차이를 보이지 못했다는 점이다. 표나 그래프에서 보여지는 것처럼 두 방식 사이에서는 눈에 띄는 성능의 차이를 찾을 수 없었다. 복잡한 계산이 필요한 KJ-WSD 방식이 KJ-Nor 방식에 비해서 눈에 띄는 성능 차이를 만들어내지 못함으로써 이 방식의 효용성에 아직은 문제가 있다고 보여진다.

각 방식별 재현율을 살펴보면 다음과 같다.

	JJ	KJ-Naive	KJ-Nor	**KJ-WSD**
T	0.6007	0.4002	0.4237	**0.4253**
D	0.6028	0.4365	**0.4434**	0.4414
N	0.7205	0.4869	**0.5747**	0.5731
C	0.6399	0.5087	0.5173	**0.5218**
TDNC	0.7453	0.5243	0.6152	**0.6166**

표2: 각 방식에 따른 재현율

KJ-Nor 방식의 평균 정확률이 전체적으로 KJ-WSD 방식에 비해서 떨어짐에도 불구하고 Description과 Narrative에서의 재현율은 오히려 더 높다. 이런 현상은 중의도가 높은 어휘의 가중치를 낮추는 효과와 더불어서 하나의 질의 어휘가 여러 대역어를 생성했을 때 이 대역어들이 유사한 의미를 가질 경우 가중치를 균일하게 부여하는 것이 관련 문서를 고루 검색할 수 있기 때문에 발생한 것으로 보인다.

6.3 중의도와 중의성 어휘 비율에 따른 성능 평가

중의성 해소 방식이 큰 성능적 우위를 점하지 못하는 데에는 질의 전체의 중의도가 관계가 있을 수 있다. 중의도가 지나치게 낮다면 가중치 부여에 의미가 없을 것이고, 반대로 중의도가 지나치게 높다면 공기 정보 계산에 있어서 올바르지 않은 대역어들 사이의 공기 정보가 섞이면서 정확성을 하락시킬 가능성이 있기 때문이다.

질의 전체의 중의도 뿐 아니고 질의의 어휘 중에서 중의성을 가지는 어휘의 비율 역시 검색의 성능에 영향을 미칠 수 있다. 여러 어휘가 2개의 의미를 가질 때와 하나의 어휘가 매우 높은 중의도를 가질 때, 질의 전체의 중의도는 비슷할 수 있지만 중의성을 해소하는 데에 있어서는 하나의 어휘가 매우 높은 중의도

를 가질 경우가 잘못된 대역어간의 공기 정보가 섞이지 않음으로써 정확한 의미를 찾아내는 데에 유리하기 때문이다.

위의 추측에 기반, 질의별로 수행된 검색 결과를 중의도와 중의성 어휘 비율에 따라 재정렬해서 새로운 결과를 얻었다.

여기서는 통계적으로 뚜렷한 의미를 가진다고 보기 힘든 결과가 도출된 Title과 Concept 필드는 제외하고 짧을 질의를 대표하는 Description과 긴 질의를 대표하는 Narrative의 결과만을 살펴보도록 하겠다.

6.3.1 Description

Description은 질의가 짧고 중의도가 다른 필드에 비해서 상당히 높다. 따라서 질의별 중의도의 편차가 심하고 중의도와 정확성 사이에 뚜렷한 상관관계도 나타나지 않았다.

각 방식별로 최고 성능과 최저 성능을 낸 질의의 개수는 다음과 같다.

	KJ-Naïve	KJ-Nor	KJ-WSD
# of Best	12	21	20
# of Worst	30	7	11

표3: 최고 성능과 최저성능 질의 개수(중복 포함)

전체적인 성능은 KJ-WSD 방식이 약간 더 높았음에도 불구하고 가장 우수한 성능을 낸 질의의 개수는 KJ-Nor 방식이 더 많았다. 특히 가장 낮은 성능을 낸 질의의 개수도 KJ-Nor이 KJ-WSD보다 적었다는 점은 주목할만한 결과이다.

KJ-WSD는 충분한 문맥 정보를 얻지 못해 중의성이 제대로 해소되지 않을 경우 최저의 성능을 나타내고 중의성이 적절히 해소될 경우에는 최고의 성능을 내는 등 성능에 편차가 심했다. 결국 몇 개의 우수한 성능을 보인 질의로 인해 전체적인 성능이 높아졌다고 볼 수 있다.

KJ-Nor이 가장 우수한 성능을 낸 질의가 많았던 것은 이는 이 방식이 잘못된 중의성 해소의 위험 없이 높은 중의도를 가지는 어휘의 가중치를 적절히 낮춰줄 수 있으며, 앞에서 언급한대로 재현율를 높이는 효과도 가지기 때문으로 보인다.

질의의 중의도와 정확률, 중의성을 가진 어휘의 비율과 정확률의 관계는 <표4>, <표5>와 같다.

KJ-Naïve 방식이 중의도가 낮은 곳에서는 다른 방식에 크게 뒤떨어지지만, 중의성이 지나치게 높아지면 다른 방식보다 우수한 성능을 보였다. 하지만 중의성을 가진 어휘 비율과 정확률의 상관관계를 보면 KJ-Naïve 방식이 거의 전구간에서 다른 방식들보다 성능이 떨어지는 것을 확인할 수 있다. KJ-Nor 방식과

중의도	KJ-Naive	KJ-Nor	KJ-WSD	# of Query
~1.10	0.3023	0.3575	0.3572	6
1.10~1.15	0.1855	0.1925	0.1921	7
1.15~1.20	0.3272	0.3528	0.3591	6
1.20~1.25	0.1033	0.1045	0.1091	2
1.25~1.30	0.1799	0.1695	0.1685	8
1.30~1.35	0.1395	0.1418	0.1385	3
1.35~1.40	0.0876	0.0657	0.0914	4
1.40~1.45	0.2084	0.2733	0.2706	8
1.45~1.50	0.2213	0.3266	0.3341	2
1.50~	0.1148	0.1038	0.0951	6

표4: Description에서의 중의도와 정확률

비율	KJ-Naive	KJ-Nor	KJ-WSD	# of Query
~0.10	0.2731	0.3256	0.3254	7
0.10~0.15	0.1855	0.1925	0.1921	7
0.15~0.20	0.2377	0.2528	0.2575	8
0.20~0.25	0.1472	0.1823	0.1783	6
0.25~0.30	0.1858	0.1970	0.2057	14
0.30~0.35	0.2404	0.2396	0.2298	3
0.35~0.40	0	0	0	0
0.40~	0.1388	0.1580	0.1530	7

표5: Description에서의 중의성을 가진 어휘 비율과 정확률

KJ-WSD 방식 사이에서는 특정한 상관관계를 찾아내기 힘들다. 구간별 성능 차이는 거의 소수의 질의에서 성능 차이가 크게 나면서 발생한 것이다.

Title과 Concept 등도 전체적으로 Description과 유사한 경향을 보였다. KJ-WSD 방식은 주로 중의도 1.30 이하, 중의성 어휘 비율이 0.15~0.30에 해당하는 구간에서 가장 우수한 성능을 나타냈고, 나머지 구간에서는 KJ-Nor 방식이 우수했으나 눈에 띌만한 차이는 발생하지 않았다.

6.3.2 Narrative

Narrative의 경우는 질의의 길이가 길고 각 질의별로 충분한 수의 중의성을 가진 어휘가 포함되기 때문에 각 가중치 부여 방식에 따른 성능의 특징이 보다 뚜렷하게 나타난다.

먼저 각 가중치 부여 방식별 최고 성능과 최저 성능을 나타낸 질의의 개수는 다음과 같다.

	KJ-Naïve	KJ-Nor	KJ-WSD
# of Best	3	37	17
# of Worst	48	5	2

표6: 최고 성능과 최저성능 질의 개수(중복 포함)

무엇보다도 KJ-Naïve 방식이 다른 방식들에 비해 확연히 뒤쳐지는 걸 알 수 있다. 또한 KJ-WSD가 최저 성능을 기록하는 경우도 크게 낮아졌다.

여전히 KJ-Nor 방식이 최고 성능을 기록하는 경우가 많았고, 여전히 KJ-WSD 방식이 평균 정확률에서 가장 우수한 성능을 보였다. 그러나 Description과는 달리 Narrative에서는 KJ-Nor 방식의 성능 편차가 KJ-WSD 방식에 비해 더 심했다. 이것은 문맥 정보가 충분히 주어진 덕에 중의성 해소가 안정적으로 작동했다고 볼 수 있다. 질의별로 봤을 때에도, KJ-Nor 방식은 KJ-Naïve 방식의 성능을 크게 밑도는 경우가 종종 발생한 데 비해서 KJ-WSD는 단 두 번을 제외하고는 항상 KJ-Naïve 방식보다는 높은 성능을 기록했다.

중의도 구간별 정확률은 다음과 같다.

중의도	KJ-Naïve	KJ-Nor	KJ-WSD	# of Query
1.10~1.15	0.3722	0.4272	0.4258	5
1.15~1.20	0.3034	0.3458	0.3426	9
1.20~1.25	0.2139	0.2960	0.3074	12
1.25~1.30	0.1477	0.2178	0.2251	16
1.30~1.35	0.1890	0.2786	0.2709	7
1.35~1.40	0.0367	0.0367	0.0329	4
1.40~1.45	0	0	0	0
1.45~1.50	0.1701	0.2540	0.2495	2

표7: Narrative에서의 중의도와 정확률

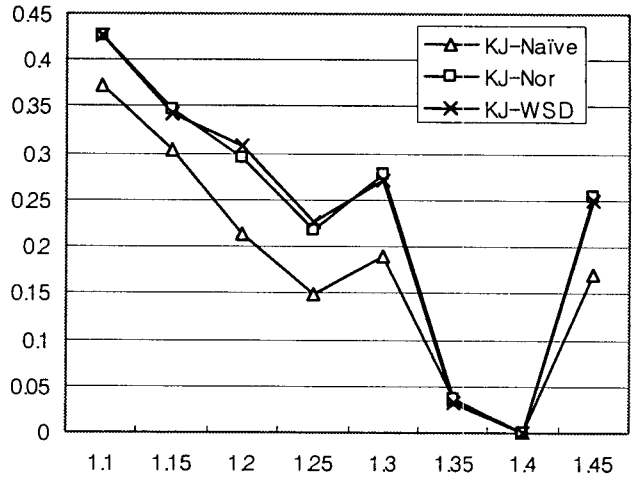
비율	KJ-Naive	KJ-Nor	KJ-WSD	# of Query
~0.10	0.2148	0.2650	0.2656	4
0.10~0.15	0.2770	0.3213	0.3198	9
0.15~0.20	0.1940	0.2751	0.2716	17
0.20~0.25	0.1999	0.2705	0.2845	18
0.25~0.30	0.2007	0.2644	0.2696	3
0.30~0.35	0.1204	0.1500	0.1352	4

표5: Narrative에서의 중의성을 가진 어휘 비율과 정확률

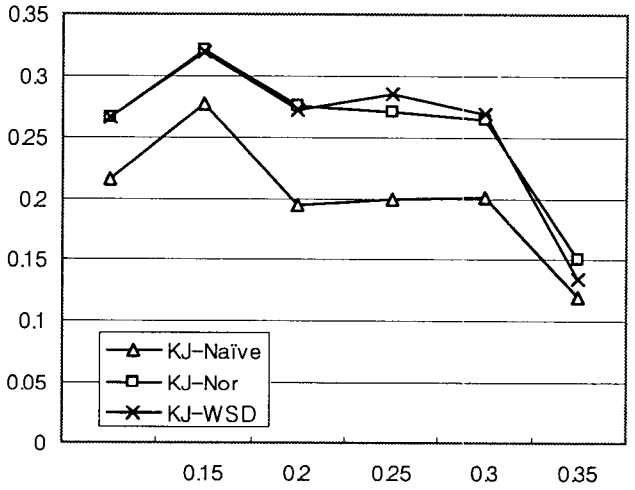
먼저 중의도와 정확률을 살펴보면 질의의 절반 이상이 중의도 1.20~1.30 구간에 속해 있었고, 이 구간에서의 평균 정확도는 KJ-WSD 방식이 가장 높았다. 그 외의 다른 구간에서는 KJ-Nor 방식의 정확률이 가장 높았다.

중의성을 가진 어휘 비율과 정확률을 살펴보면 상관관계가 더욱 두드러진다. KJ-Naïve 방식은 가장 뚜렷하게 최저 성능을 그리고 있고, KJ-WSD 방식은 0.20~0.30 구간에서 다른 구간을 월등히 뛰어넘는 성능을 보이고 있고 그 외 다른 구간에서는 KJ-Nor의 성능이 더 우수했다.

그래프3: Narrative에서의 중의도와 정확률



그래프4: Narrative에서의 중의성을 가진 어휘 비율과 정확률



전반적으로 중의도나 중의성을 가진 어휘 비율이 너무 높거나 낮지 않은 상황에서 KJ-Nor 방식과 KJ-WSD 방식의 성능이 KJ-Naïve 방식에 비해서 높아진다는 점은 짧은 질의에서 확인한 결과와 대체로 일치하고 있다. 또한 질의가 길어질 경우에는 KJ-WSD 방식이 좀 더 확실한 우수성을 보이는 구간이 나타나는데, 이것은 질의의 길이가 KJ-WSD의 성능에 영향을 미치는 증거라고 할 수 있다. 즉 어느 정도 이상의 길이의 질의가 주어졌을 때, 중의도와 중의성을 가진 어휘의 비율이 적정 수준에 들어왔다면 KJ-WSD 방식이 잘 작동한다고 볼 수 있는 것이다. 이 가정을 토대로 휴리스틱을 만들어서 실험한 결과, KJ-WSD 방식에 비해 공기 정보를 계산하는 양을 절반으로 줄이면서 1% 안팎의 미묘한 성능 향상 효과까지 얻었다.

그러나 이 휴리스틱은 다른 문서 집합에서도 효율적인지 실험을 통해 확인할 필요가 있겠다.

7. 결론

전체적으로 중의도나 중의성을 가진 어휘의 비율은 가중치 부여방식의 성능에 약간의 영향을 미치는 것으로 보인다. 그리고 질의의 길이가 길수록 가중치를 부여하는 방식의 성능이 더욱 크게 올라가는 반면에 질의의 길이가 짧을 경우에는 가중치 부여 방식 자체가 의미가 없어지게 된다.

중의도나 중의성 어휘 비율이 너무 낮을 경우에는 대역어에 가중치를 부여하는 것과 부여하지 않는 것이 큰 차이를 유발하지 않기 때문에 가중치 부여 방식에 상관없이 성능이 유사한 점은 당연하다고 볼 수 있다.

반면에 중의도나 중의성 어휘 비율이 너무 높을 경우 성능 향상폭이 오히려 조금씩 줄어드는 현상은 KJ-Nor 방식과 KJ-WSD 방식의 경우를 각각 나눠서 생각해볼 수 있다.

KJ-Nor 방식에서는 질의의 중의도가 높을 경우 대역어의 가중치를 낮춤으로써 잘못된 문서가 검색될 확률을 낮춰주지만, 역효과로 올바른 대역어에조차 너무 낮은 가중치가 부여됨으로써 올바른 문서를 검색하는데에 지장을 초래하게 된다. 중요한 어휘가 높은 중의도로 낮은 가중치를 받고, 검색에 도움이 안되는 어휘가 낮은 중의도로 높은 가중치를 받는다면 올바른 검색이 이루어질 수 없게 되는 것이다.

KJ-WSD 방식에서는 질의의 중의도나 중의성이 있는 어휘의 비율이 높아질 경우, 잘못된 대역어들과의 공기 정보의 비중이 높아지면서 생기는 잡음 효과로 인해서 정확한 중의성 해소를 방해하게 된다. KJ-WSD 방식의 이러한 문제점은 상대적으로 질의의 중의도나 중의성이 있는 어휘의 비율에 덜 민감한 KJ-Nor 방식을 적용함으로써 약간의 보완은 가능했지만, 그것만으로 눈에 띄는 차이를 가져오기에는 역부족이었다.

본 논문에서는 통계적으로 대역어의 변환 확률을 계산하는 모델을 제시했다. 그러나 이 모델을 적용해서 중의성을 해소하려고 했던 KJ-WSD 방식은 현재까지의 실험으로는 다른 통계적 모델이 그러했던 것처럼 상대적으로 큰 계산 복잡성에도 불구하고, 단순히 질의 어휘의 중의성에 반비례하는 가중치를 부여하는 방식인 KJ-Nor 방식의 성능을 뛰어넘었다고 볼 수 없는 수준의 성능밖에 보여주지 못하였다. 따라서 이 모델은 단점을 분석하고 이를 보완, 더 높은 성능을 낼 수 있도록 하기 위한 연구가 지속되어야 할 것으로 보인다.

8. 향후 연구

처음에 제시했던 대역어 확률 모델을 이용한 가중치 부여 방식에 있어서 가장 큰 문제점은 중의도나 중의성이 있는 어휘의 비율이 높아질수록 성능 향상의 폭이 더욱 커지는 대신 오히려 줄어드는 데 있다고 볼 수 있다. 따라서 이 방식의 성능을 높이기 위해서는 중의도와 중의성을 가진 어휘 비율이 높은 질의를 대상으로도 높은 성능을 낼 수 있도록 만드는 데에 있다고 볼 수 있다.

이를 위해서는 현재처럼 모든 문맥정보를 이용하는 대신에 밀접한 관계에 있는 어휘로부터만 공기 정보를 구해서 중의성을 해소하는 방식을 적용해볼 필요가 있다.

질의가 들어왔을 때 [5]와 유사한 방법으로 질의와 동일한 언어의 문서 색인을 이용해서 질의에 나타난 어휘를 노드로 가지는 의존 트리를 생성, 트리로 연결된 대역어들에 한해서 공기 정보를 계산하는 방식을 실험할 계획이다. 이 방식은 질의와 동일한 언어의 문서 색인과 질의어휘들의 공기 정보를 계산해야 하는 필요성이 있는 대신에, 신장 트리로 연결된 노드만의 공기 정보를 계산하기 때문에 계산의 양이 줄고 관계 없는 대역어들 사이의 공기 정보가 가중치에 반영되는 것을 최소화하며, 질의의 중의도가 1.5에도 채 미치지 않는 특성상 질의 전체의 중의도 변동에도 지금보다 덜 민감하게 반응할 것으로 기대된다.

그리고 [8]의 연구와 유사한 방법으로 현재의 모델을 가지고 가상 피드백 단계에서 EM 알고리즘을 통해서 가중치를 조절하는 방식을 적용하는 연구를 적용할 예정이다.

참고 문헌

[1] L. Ballesteros, and W. B. Croft. Resolving ambiguity for cross-language retrieval. In Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 64-71. 1998.

[2] K. Darwish, and D. W. Oard. Probabilistic structured query methods. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 338-344. 2003.

[3] M. Federico, and N. Bertoldi. Statistical cross-language information retrieval using N-best query translation. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 167-174. 2002.

[4] J. Gao, J. Nie, H. He, W. Chen, and M. Zhou. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,

pages 183-190. 2002.

[5] J. Gao, J. Nie, G. Wu, and G. Cao. Dependence Language Model for Information Retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 170-177. 2004.

[6] J. Gao, J. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving query translation for cross-language information retrieval using statistical models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 96-104. 2001.

[7] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-Lingual Relevance Models. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 175-182. 2002.

[8] C. Monz, and B. J. Dorr. Iterative translation disambiguation for cross-language information retrieval. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 520-527. 2005.

[9] D. W. Oard, and B. J. Door. A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. 1996.

[10] A. Pirkola. The effects of query structure and dictionary setups in dictionary based cross-language information retrieval. In Proceedings of the 21th Annual International SIGIR Conference on Research and Development in Information Retrieval, pages 55-63. 1998.

[11] A. Pirkola, T. HedLund, H. Keskustalo, and K. Järvelin. Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 4, Kluwer Academic Publishers, pages 209-230. 2001.

[12] P. Sheridan, and J. Ballerini. Experiments in multilingual information retrieval using SPIDER system. In Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pages 58-65. 1996.

[13] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 105-110. 2001.

[14] 이규찬, 강인수, 나승훈, 이종혁. 한국컴퓨터 종합학술대회 2005 논문집(B), pages 448-450.