바이그램 색인에 기반한 한-일 교차언어검색

이규찬°, 강인수, 나승훈, 이종혁
포항공과대학교 지식 및 언어공학 연구실
{l17690°, dbaisk, nsh1979, jhlee}@postech.ac.kr

Korean-Japanese Cross Lingual Information Retrieval Based on Bi-gram Indexing

Gyu-Chan Lee°, In-Su Kang, Seung-Hoon Na, Jong-Hyeok Lee
Knowledge & Language Engineering Lab, Pohang University of Science and Technology

요 약

교차언어검색 시스템은 다양한 언어자원을 필요로 한다. 여기서는 한-일 대역어 사전과 일본어 문서의 바이그램 색인만을 이용해서 교차언어검색을 수행하는 방법을 제시한다. 한국어로 된 자연어 질의에서 형태소 분석기 등의 도움 없이 간단하게 일본어 대역어 리스트를 생성할 수 있는 방법과, 검색의 성능을 올릴 수 있도록 대역어에 가중치를 부여하는 방법을 제안한다. 그리고 실험을 통해 제시한 방법을 평가하고 분석한다.

1. 서론

교차언어검색은 질의와 문서의 언어가 서로 다른 상황에서의 검색을 의미한다[1]. 따라서 검색을 수행하기에 앞서 질의 언어와 문서의 언어를 서로 일치, 비교 가능하게 하는 변환 과정이 필요하다.

변환의 대상을 기준으로 보았을 경우, 질의를 변환하는 방식과 문서 전체를 변환하는 방식 두 가지가 있으며, 변환을 위한 언어 자원으로는 기계번역 시스템, 병렬 코퍼스, 혹은 대역어 사전 등이 사용된다[2]. 언어 자원과 시스템 구축의 용이성, 시스템 전체의 오버헤드 측면에서 볼 때 다른 방법에 비해서 사전을 이용한 질의 변환 방식이 선호된다.

이 논문에서는 일본어 문서의 바이그램 색인과 사전을 이용하여 교차언어검색을 수행하는 방식, 그리고 바이그램 색인에서 효과적인 가중치 부여 방식에 대해서 논하고자 한다.

한국어와 일본어의 경우 띄어쓰기가 없거나 어절과 어휘가 1:1로 대응되지 않으며, 따라서 정확한 어휘 추출을 위해서는 높은 수준의 형태소 분석기가 요구된다. 그러나 형태소 분석기의 구축, 실행을 위한 오버헤드가 발생하고, 형태소 분석기의 성능이 뛰어나지 않으면 검색 성능에 악영향을 주게 된다.

색인 과정에서 형태소 분석을 피하기 위해서는 주로 바이그램 색인이 많이 사용된다. 바이그램 색인은 과정이 간단하고, 오류나 오타에 민감하지 않으며, 어휘 기반 색인에 비해 성능이 크게 떨어지지 않는다는 장점을 가지고 있으며 동아시아 언어를 처리하는 데에 선호되는 경향이 있다[3].

사전을 사용해서 질의를 변환하는 과정에서는 중의성이 발생할 수 있으며, 이로 인해 사용자의 의도와 다른 의미를 가진 어휘가 대역어에 포함될 경우 검색의 성능을 저하시키는 요인이 된다. 따라서 성능을 높이기 위해서는 중의성을 해소하기 위한 모듈이 필요하다[2, 4, 5, 6].

이 논문에서는 먼저 중의성을 해소하기 위해서 기존의 연구에서 수행되었던 방식들의 특징과 한계점에 대해서 분석한다.

그리고 질의에서 형태소 분석기 없이 어휘를 추출하고 대역어를 생성하는 방식과, 생성된 대역어들을 그룹으로 묶어서 가중치를 부여하는 방식을 제안한다.

마지막으로 실험을 통해서 일본어 단일어 검색, 수작업으로 대역어를 추출한 경우와 비교, 이 논문에서 제시한 방법의 유용성을 판단하고 장점과 단점에 대해서 분석할 것이다.

2. 기존의 연구

교차언어검색이 단일어검색과 다른 점은 질의와 문서의 언어가 다르다는 점이고, 이의 해결을 위해 언어를 변환하는 과정에서 생기는 대역어 중의성 문제는 교차언어검색의 성능이 단일어 검색의 성능보다 떨어지게 하는 중요한 요인이다.

기계번역기를 이용해서 변환을 수행할 경우, 변환 중에 생기는 문제는 번역기 자체에 있는 모듈에서 해결하기 때문에 특별히 다른 과정이 필요하지 않은 장점이 있다[8]. 그러나 기계번역기의 구축과 실행을 위한 오버헤드가 크고, 문서 특징에 맞도록 번역기를 수정하는 것이 매우 힘들며, 번역기의 성능이 아직 신뢰할만한 수준에 이르지 못했다는 문제가 있다.

병렬 코퍼스를 이용하는 방식 역시 각 정렬의 확률값을 이용해서 대역어를 결정하므로 특별히 중의성 해소를 위한 모듈을 필요로 하지 않으며, 성능 역시 단일어 검색과 거의 비슷한 수준이다[7]. 그러나 충분한 양의 병렬 코퍼스를 구하거나 제작하는 것, 그리고 병렬 코퍼스를 서로 매칭하는 문장, 단어로 정렬하는 문제가 쉽지 않다는 문제점이 있다.

사전을 이용하는 방식은 상대적으로 위의 방식들보다 간단한 대신 사전을 통과하는 과정에서 필연적으로 발생하는 중의성 문제를 해결하는 것이 성능을 높이기 위한 관건이 된다.

Pirkola는 98년, 핀란드-영어 교차언어 검색에서 변환된 질의어들을 그룹으로 묶어서 검색을 수행하는 방식을 제안했다[2, 4]. [2]에서는 원래 질의어의 같은 단어에서 나온 대역어들과 합성어의 분해를 통해 얻은 대역어들을 그룹으로 묶어서 가중치를 조절할 경우, 단순히 대역어들의 나열을 이용하는 것보다 30% ~50% 정도 성능이 향상되었다고 보고하고 있으며, 이는 단일어 검색 성능의 80% 수준이다. [4]에서는 다른 언어 쌍에 대해서 위의 실험을 확대, 모든 언어 쌍에서 16%~43% 정도 성능이 향상되었음을 보이고 있다.

단순한 구조화보다 적극적으로 중의성을 해소하는 방식으로 가장 많이 쓰이는 방식은 통계적 방식을 이용하는 것이다[5, 6]. 이 방식은 어떤 단어가 번역 과정에서 중의성을 가질 경우, 올바른 번역어들은 그렇지 않은 번역어들에 비해서 주변의 다른 대역어들과 서로 공기할 가능성이 높다는 가정에 기반한다. 이 방식들은 검색 대상 문서 집합에서 공기 정보를 뽑아내는 데 상당한 시간과 메모리를 필요로 한다는 문제점에도 불구하고, 그 성능은 단일어 검색 성능의 80% 전후 수준에 그쳐, 단순히 질의를 구조화하는 방식을 크게 능가하지는 못했다.

또 다른 방식으로는 구 단위로 변환을 수행하는 방법이 제안되었다[7]. 구는 어휘보다 훨씬 구체적인 정보를 가지므로 보다 낮은 중의성을 가진다. 그러나 이를 위해서는 높은 정확성을 가지고 구를 추출해내는 기술이 필요하고, 구를 추출하는 과정에서 상당한 자원의 소모가 불가피해진다. 또한 [7]의 방식은 구 추출을 위해서 중국어 의존적인 특성을 이용하기 때문에 다른 언어간에서는 그대로 사용할 수 없다는 문제점을 지니고 있다.

3. 질의 변환

한국어는 일본어 문서 색인에서와 마찬가지로 어휘와 어절이 1:1로 대응되지 않기 때문에 질의에서 어휘를 추출하는 과정을 필요로 한다.

형태소 분석기를 사용할 경우, 어휘를 뽑아내고 용언의 원형을 복원해줄 수 있다는 장점이 있지만, 형태소 분석기 구축과 실행을 위한 시간과 자원을 필요로 하고, 검색 성능이 분석기의 성능에 크게 영향을 받게 되는 단점도 있다. 특히 질의에서 중요 어휘가 누락되면 검색에 치명적인 악영향을 주게 된다.

누락되는 어휘 없이 간단하게 어휘를 추출할 수 있는 방법으로 모든 길이의 n-그램을 이용하는 방법이 있다. 이 방법은 하나의 어절에 대해 모든 가능한 부분문자열을 추출하는 것으로, 예를 들어 "대학생"이라는 어절이 있다면 {대, 학, 생, 대학, 학생, 대학생}의 6개 문자열을 추출하는 방법이다.

이 방식은 {대, 학, 생} 같이, 검색에 영향을 주지 않거나 잡음으로 작용하는 문자열을 뽑아낼 수도 있고, 특히 잘못된 1음절 단어가 뽑혀져 나올 가능성이 상당히 높다. 또한 동사의 원형을 복원할 수 없다는 단점을 가진다.

그러나, 일반적으로 짧은 길이의 질의에서는 체언이 용언보다 중요한 경우가 많아 체언을 빠짐없이 추출해내는 것이 더 중요하며, 복합 명사를 분해해서 관련 어휘를 추가하거나 강조하는 효과를 가질 수 있다. 위의 경우에도 {대학생}이라는 올바른 어휘 외에 {대학, 학생} 같이 관련된 어휘를 뽑아냄으로써 관련 문서를 찾아내는 데에 도움을 주는 효과를 기대할 수 있다.

모든 길이의 n-그램을 통해 얻은 문자열들을 한-일 사전에 통과시키면 대역어를 얻을 수 있다. 그러나 이 대역어들은 많은 오류 단어들을 포함하고 있다. 특히 "한"이나 "제" 같은 출현 빈도가 높고, 높은 중의성을 가지는 어휘로 잘못 인식될 수 있는 음절들이 검색 성능을 크게 하락시킬 수 있다는 점을 1음절 어휘를 제거한 경우의 실험 결과를 통해서 확인할 수 있다.

4. 가중치 조절

원래 질의에서 추출된 문자열들은 다양한 개수의 대역어를 가질 수 있다. 따라서, 모든 대역어들을 같은 중요도로 보는 것보다는, 원어의 중의성을 고려해서 가중치를 부여하면 보다 정확한 검색 결과를 얻을 수 있을 것이다.

가장 먼저, 한 개의 어휘에서 n개의 대역어가 나왔을 경우, 각각의 대역어에 1/n의 가중치를 부여하는 방식을 생각할 수 있다. 중의성이 심한 어휘로부터 얻어진 대역어일수록 낮은 점수를 주어서 검색에 영향력을 제한하는 방식이다. 이 방식은 언어쌍이 다르고 문서가 바이그램으로 색인되어 있다는 점을 제외하면 [2]의 방식과 같은 방식이라고 볼 수 있다. 이 가중치 부여 방식을 N1이라고 한다.

그러나 바이그램 색인에서 검색을 수행하기 위해 검색어의 바이그램을 추출할 경우, 길이가 긴 대역어에서 추출되는 바이그램의 개수가 검색어가 짧은 대역어에서 추출되는 바이그램의 개수보다 많기 때문에, N1은 길이가 긴 대역어에 유리하다고 볼 수 있다. 대역어 길이 요소를 제거하기 위한 방법으로 N1에 대역어 텀이 생성한 바이그램의 개수로 다시 나눠준 값을 가중치로 생각해볼 수 있다. 즉 하나의 어휘가 3개의 대역어를 생성했고, 그 중 한 대역어가 3음절로 이루어졌다면 그 대역어는 2개의 바이그램을 생성하므로 그 대역어의 바이그램의 가중치는 각각 1/(3*2)가 된다. 이 방식은 N2로 지칭한다.

N2의 가정과는 반대로, 길이가 긴 어휘가 길이가 짧은 어휘에 비해서 검색을 수행하는 데에 결정적인 역할을 한다고 볼 수도 있다. 즉, 2음절 어휘와 4음절 어휘가 있다면 4음절 쪽이 의미가 더 명확하거나 더 구체적일 수 있다는 것이다. N1은 이 가정은 만족하지만 다른 문제를 가진다. 하나의 어휘에서 생성된 대역어들의 바이그램의 점수의 합이 전부 일정하지 않는다는 점이다. 어떤 어휘가 3개의 대역어를 가지며, 이 때 각 대역어들이 가지는 바이그램의 개수의 합이 4개라면 이 바이그램들의 점수의 합은 (1/3)*4 = 4/3 이 된다. 반면에, 어떤 어휘가 1개의 대역어만을 가지며 그 대역어가 4개의 바이그램을 생성한다면, 이 때 점수의 합은 1*4 = 4 가 된다. 즉, 이 경우에는 원래 질의의 두 어휘 중 두 번째 어휘가 더 중요한 것처럼 취급된다고 볼 수 있다. 이런 문제를 해결하기 위해서, N1에서 사용된 대역어의 개수 대신 전체 바이그램의 개수로 가중치를 부여함으로써 하나의 질의 어휘에서 나온 대역어의 바이그램의 점수의 합을 1로 만들 수 있다. 이 방식을 N3라고 지칭한다.

각각의 실험 결과는 다음 장에서 설명하도록 한다.

5. 실험 결과 및 평가

5.1 실험 환경

실험을 위한 문서로는 NTCIR-3 일본어 문서 집합을 사용하였다. 이 문서 집합은 1998~1999년 2년 간의 일본의 마이니치 신문의 기사 220078건으로 구성되어 있으며, 바이그램 방식으로 색인되었다.

질의는 NTCIR-3에서 제공된 98년 문서용 질의 50건 중 일본어 문서 검색을 위해 선택된 41개의 질의를 사용하였다. 이 질의는 전문가에 의해서 한, 중, 일, 영 4가지 언어로 되어 있으며, 여기서는 일본어와 한국어 질의를 사용하였다.

질의 변환을 위한 한-일 사전은 포항공대 지식 및 언어공학 연구실에서 개발한 Cobalt 한-일 번역기에서 사용되는 사전을 변환해서 사용하였다.

검색 순위 함수는 Okapi-BM25모델을 사용했다.

5.2 실험 및 결과

성능의 평가 수단으로는 평균 정확률과 R-정확률을 이용했다. 평균 정확률은 올바른 문서가 검색된 순간의 정확률의 평균값을 의미하며, R-정확률은 검색된 문서들을 상위에서부터 질의와 관련있는 문서의 개수만큼 뽑아냈을 때, 뽑혀진 문서 내에서의 정확률을 의미한다.

JJ는 일본어 단일어검색을 의미한다. 교차언어검색은 언어 변환 오류 등으로 인해서 단일어검색 성능을 넘기 힘들다는 점에서 교차언어검색 성능의 상한선이라고 할 수 있다.

KJ-1은 얻어진 대역어를 아무 가공 없이 그대로 사용한 가장

단순한 방법으로, 이 실험의 기본 모델이라고 볼 수 있다.

KJ-2는 KJ-1에서 쓰인 질의에서 한국어와 일본어를 모두 아는 사람에 의해서 올바른 대역어만 남기고 검색을 수행한 것으로, 현재 사전을 통해 낼 수 있는 최대 성능이라고 볼 수 있다.

KJ-N1부터 KJ-N3는 각각 위에서 언급한 N1, N2, N3 가중치를 이용해서 실험을 수행했을 때 나타난 결과이다.

KJ-WSD는 공기 정보를 이용해서 원래 질의에 나타난 어휘가 각각의 대역어로 변환될 확률을 계산한 방식으로, 아직 실험이 미흡하여 자세한 설명은 생략한다.

<표1>은 위 실험의 결과를 나타낸 것이다.

<표1> 가중치 할당 방식에 따른 성능 비교

검색 방식	평균 정확률	R-정확률
JJ	0.2770	0.2985
KJ-1	0.1847	0.2155
KJ-2	0.2516	0.2855
KJ-N1	0.2308	0.2565
KJ-N2	0.1906	0.2168
KJ-N3	0.1916	0.2204
KJ-WSD	0.2277	0.2522

결과에 의하면 N1이 단일어 검색 성능의 83.32%(0.2308/0.2770)에 달하는 평균 정확률을 기록, 가장 뛰어난 성능을 보였다. 이는 사전을 이용한 경우의 한계치인 KJ 2의 정확률의 91.73%(0.2308/0.2516)에 이르며, 사실상 거의 한계에 가까운 성능을 낸 것으로 볼 수 있다.

WSD의 평균 정확률은 N1의 평균 정확률보다도 오히려 약간 낮은 수준에 그치고 있다.

<표2>는 모든 길이의 n-그램을 뽑는 과정에서 1음절 어휘를 제거한 후 한-일 사전을 통과시켜서 대역어를 생성해서 KJ-1과 KJ-Nn 실험을 반복한 것이다.

<표2> 1음절 어휘를 제거한 후 대역어를 생성했을 때

검색 방식	평균 정확률	R-정확률
KJ-1	0.1993	0.2266
KJ-N1	0.2396	0.2622
KJ-N2	0.2156	0.2387
KJ-N3	0.2169	0.2381
KJ-WSD	0.2368	0.2600

1음절 어휘를 제거한 결과에서도 역시 N1의 성능이 가장 우수했다. N1과 WSD 성능이 거의 같이 뛰고, N2와 N3의 성능이 다른 경우에 비해서 상당히 크게 뛴 점이 특징이다.

6. 분석 및 토의

실험 결과에 의하면 어휘의 대역어 개수를 이용해서 가중치를 부여한 N1 방식이 가장 우수한 성능을 보였다. 특히 바이그램의 개수를 고려한 N2나 N3 방식보다도 우수한 성능으로, 이는 바이그램 색인을 만들 때, 같은 어휘에서 나온 바이그램의 개수를 고려하지 않은 것처럼 질의에서도 같은 어휘에서 나온 바이그램을 고려하지 않음으로써 질의와 색인 구조를 똑같이 유지하게 하는 것이 바람직하다고 볼 수 있고, 대역어의 길이나, 하나의 질의어휘에서 나온 대역어들의 가중치의 합을 고려하는 것은 옳지 않다고 볼 수 있겠다.

WSD를 수행한 결과는, 다른 기존 연구에서 제시한 방식과 비슷한 성능을 보였음에도 불구하고 KJ-N1의 성능을 뛰어넘지 못하였다. 이는 중의성 해소가 단지 잘못된 대역어의 가중치를 상대적으로 낮추는 효과만 가지고 있을 뿐, 정확한 중의성 해소가 되지 못했다는 의미일 수 있다. 혹은 실제 질의에서 성능에 큰 영향을 미칠 만큼 중의성을 가진 어휘가 많지 않다는 점을 의미할 수도 있다. 확실한 결론을 위해서는 더욱 큰 문서집합에서 실험을 수행, 결과를 확인할 필요가 있을 것이다.

7. 결론

일본어 문서의 바이그램 색인과 한-일 사전 두 가지 자원을 이용해서 교차언어 검색을 수행하는 시스템을 구현해 보았다.

한국어 단일어검색, 일본어 단일어검색 모두에서 효과적이라고 보고된 bi-gram 색인이 한-일 교차언어검색에 있어서도 충분한 성능을 보인다는 점을 확인할 수 있었다.

특히, 형태소 분석기 등 부가적인 시스템의 도움을 배제하여 오버헤드를 최소화하면서도 교차언어검색에서 단일어검색의 평균 정확률의 80% 수준의 성능을 낼 수 있다는 점이 이 실험에서 가장 큰 의의라고 할 수 있다.

또한 다양한 가중치 부여 방식 중 단순히 대역어들을 그룹으로 묶어서 가중치를 부여하는 방식이, 복잡한 수식을 통한 중의성 해소 방식보다도 더 뛰어난 성능을 보였고, 그 성능이 한-일 사전을 이용한 경우의 성능 한계의 90%에 달했다는 점도 주목할만한 결과이다.

참고 문헌

[1] Douglas W. Oard & Bonnie J. Door, 1996, " A Survey of Multilingual Text Retrieval" Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies

[2] Ari Pirkola, 1998, " The Effects of Query Structure and Dictionary Setups in Dictionary Based Cross-language Information Retrieval" In Proceedings of the 21th SIGIR Conference on Research and Development in Information Retrieval, pages 55-63

[3] Ogawa Yasushi, Kameda Masayuki & Matsuda Toru, 1996, " Inforium: A User-friendly Document Retrieval System" In Proceedings of the workshop on Information Retrieval with Oriental Languages, pages 143-149

[4] Ari Pirkola, Turid HedLund, Heikki Keskustalo & Kalervo Järvelin, 2001, " Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings" Information Retrieval, 4, pages 209-230, Kluwer Academic Publishers

[5] Lisa Ballesteros & W. Bruce Croft, 1998, " Resolving Ambiguity for Cross-language Retrieval" , In Proceedings of the 21th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pages 64-71

[6] Jianfeng Gao, Jian-Yun Nie, Hongzhao He, Weijun Chen & Ming Zhou, 2002, " Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations" , In Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pages 183-190

[7] Sheridan, P. & Ballerini, J. 1996. " 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