# 병렬말뭉치를 이용한 대체어 자동 추출 방법

백종범○ 이수원
숭실대학교 대학원 컴퓨터학과
jbb100@ssu.ac.kr, swlee@ssu.ac.kr

# Automatic Extraction of Alternative Words using Parallel Corpus

Jongbum Baik○ Soowon Lee
Dept. of Computing, Graduate School, Soongsil University

## 1. 서 론

키워드 기반의 정보검색에서 사용자가 원하는 정보가 누락되는 현상은 주로 사용자가 적합한 키워드를 선정하지 못함으로 인하여 발생한다. 이러한 키워드 선정의 어려움은 '어휘 표기의 다양성'으로부터 시작된다. 예를 들어 사용자가 'Television'과 관련된 문헌을 검색하는 경우에 정보 누락을 최소화하기 위해서는 '텔레비전', '텔레비젼', '텔레비죤', '테레비전' 등 다양한 표기 형식을 고려한 질의문(Query)을 작성해야 한다. 그러나 현실적으로 사용자가 위와 같은 다양한 형태의 표기 형식을 모두 유추하여 질의문을 작성하는 것은 불가능하며, 비록 가능하다 할지라도 이는 사용자에게 수많은 시간과 노력을 소모하게 만들 것이다. 본 연구에서는 이와 같은 '어휘 표기의 다양성'으로 인한 정보 누락을 최소화하기 위하여 병렬말뭉치를 이용하여 대체어를 자동으로 추출하는 방법을 제안한다.

본 연구에서 정의하는 대체어란, "한 문장에서 특정 단어를 대신하여 사용해도 문장의 의미를 훼손하지 않는 단어"를 의미한다. 본 연구에서는 대체어를 이형어, 대역어, 유의어로 분류한다[1]. 특히, 본 연구에서는 3가지 대체어 유형 중에서 사용자가 직접 유추하기 힘든 '이형어'와 교차언어검색에 활용할 수 있는 '대역어'를 추출하는 것에 중점을 둔다.

대체어를 자동으로 추출하기 위한 대부분의 연구들은 특정 단어 주변의 문맥(Context) 정보를 이용하여 대체어를 추출한다[1-4]. 이러한 연구들은 대체어일 가능성이 높은 단어들을 추출하는 데에는 많은 공헌을 하였으나, 최종적으로 대체어 목록을 자동으로 결정하기 위한 '대체어 결정함수'로 발전시키지 못하였다는 점에 있어서 한계를 지닌다. 이러한 문제는 단어 간 동시출현빈도에 기반한 연관단어 뭉치를 각 단어의 특징(Feature)으로 이용하여 대체어를 추출함으로 인하여 발생한다. 왜냐하면 연관단어 뭉치를 각 단어의 특징으로 이용할 경우, 특징 선택(Feature Selection) 기준을 명확하게 정의하는 것이 쉽지 않기 때문이다. 본 연구에서는 이러한 연관단어 뭉치의 단점을 극복하기 위하여 병렬말뭉치로부터 추출한 대역어 뭉치를 각 단어의 특징으로 이용한다.

## 2. 대체어 자동 추출 시스템

본 연구에서 제안하는 대체어 자동 추출 시스템은 먼저 국/영문 제목(병렬말뭉치)을 이용하여 연관단어 뭉치 및 대역어 뭉치를 추출한다. 그 다음, 각 단어별 대역어 뭉치 간의 유사도를 비교하여 대체어 목록을 생성하고, 마지막으로 연관단어 뭉치를 이용하여 대체어 목록을 필터링한다.

### 2.1 단어 간 상관성 분석

본 단계에서는 국/영문 제목 간 '한글-영어 단어 쌍'의 동시출현정보를 이용하여 '대역어 뭉치'를 추출하고, 국문 제목 내 출현 단어 간 동시출현정보를 이용하여 '연관단어 뭉치'를 추출한다. 본 단계에서 추출하는 대역어 뭉치는 대체어 추출 단계(2.2절)에서 이용되며, 연관단어 뭉치는 연관단어 필터링 단계(2.3절)에서 이용된다.

본 연구에서는 대역어 뭉치를 추출하기 위해서 국문 제목과 영문 제목 간에 동시 출현한 '한글-영어 단어 쌍'의 빈도를 이용하여 Jaccard 상관계수를 계산한다[5]. 이는 기준단어와 함께 가장 많이 출현하는 대역어를 찾기 위한 과정이다. 대역어 뭉치 추출 과정에 있어서 $w_1$은 국문 제목 내에 출현한 '한글 단어'로 정의하고, $w_2$는 영문 제목 내에 출현한 '영어 단어'로 정의한다.

또한 연관단어 뭉치를 추출하는 과정에 있어서는 국문 제목 내 출현 단어 간 PMI(Pointwise Mutual Information)[2]를 계산한다. 연관단어 뭉치 추출 과정에 있어서 $w_1$, $w_2$는 모두 국문 제목 내에 출현한 '한글 단어'로 정의한다.

$$Jaccard(w_1,w_2)=\frac{|w_1\cap w_2|}{|w_1\cup w_2|} \qquad PMI(w_1,w_2)=\log\frac{p(w_1\cap w_2)}{P(w_1)p(w_2)}$$

### 2.2 대체어 추출

본 연구에서는 대체어 목록을 추출하기 위하여 기준단어와 조합 가능한 모든 단어 간의 코사인 유사도(Cosine Similarity)[5]를 계산한다. 두 단어 간의 코사인 유사도를 계산하기 위해서는 각 단어를 설명하는 특징벡터가 존재해야 한다. 본 연구에서는 2.1절에서 추출한 대역어 뭉치를 각 단어의 특징벡터로 정의하여 단어 간의 유사도를 계산한다.

### 2.3 연관단어 필터링

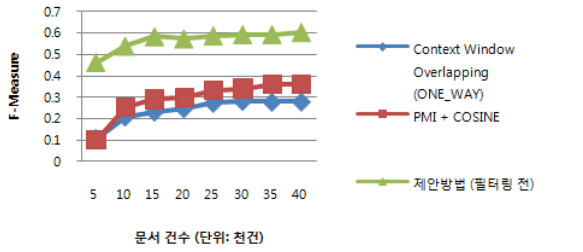
코사인 유사도를 이용하여 추출한 대체어 목록 내에는 대체어가 아닌 연관단어들이 다수 포함되는 문제가 발생한다. 이는 각 단어의 특징벡터로 이용하는 대역어 뭉치의 품질이 완전하지 못하기 때문에 발생하는 문제이다. 본 연구에서는 이러한 문제로 인한 대체어 목록 품질 저하를 최소화하기 위하여 2.1절에서 추출한 연관단어 뭉치를 이용한 연관단어 필터링 방법을 제안한다. 본 단계의 기본 아이디어는 [1]에서 제안한 "대체어는 동일 제목 내에서 출현할 확률이 적을 것이다"라는 가설에 근거한다.

PMI(수식)는 상관성 척도로서 0을 지닐 경우에는 두 단어가 독립적이라고 판단하며, 0보다 클 때에는 양의 상관관계, 0보다 작을 때에는 음의 상관 관계를 지닌 것으로 판단한다[2]. 본 연구에서는 이와 같은 PMI의 수식적 의미를 이용하여 이전 단계에서 추출한 대체어 목록 내 단어 간의 PMI를 계산한 후, 독립적(PMI=0) 혹은 음의 상관 관계(PMI > 0)를 지니는 단어 쌍만을 취하여 대체어 목록으로 결정한다.

## 3. 실험 및 결론

본 연구에서는 한국특허정보원에서 운영하는 특허정보검색서비스인 KIPRIS(http://www.kipris.or.kr/) 내의 특허정보 중 '화상 통신(H04N)', '전기에 의한 디지털 데이터 처리(G06F)' 분류로부터 각각 46,059건 및 54,669건의 국/영문 제목을 수집하여 실험을 수행하였다. 평가는 사전에 구축된 평가지표를 이용하여 MAP(Mean Average Precision)[5]와 Recall을 이용하였다. 또한 MAP와 Recall을 종합적으로 고려한 성능을 평가하기 위하여 F-Measure를 계산함으로써 시스템의 최종적인 성능을 평가하였다.

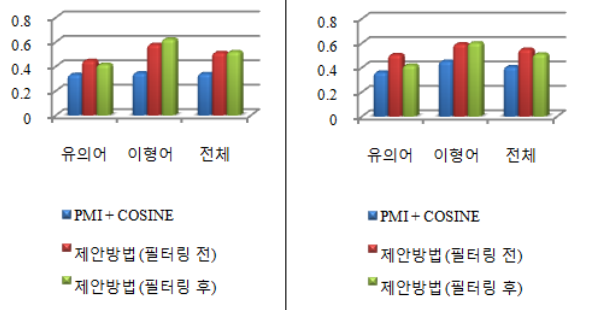
| 문서 건수 (단위:천건) | CWO (ONE_WAY) | PMI + COSINE | 제안방법 (필터링 전) |
|---|---|---|---|
| 5 | 0.1095 | 0.1033 | 0.4627 |
| 10 | 0.2102 | 0.2543 | 0.5396 |
| 15 | 0.2336 | 0.2904 | 0.5847 |
| 20 | 0.2497 | 0.2985 | 0.5739 |
| 25 | 0.2746 | 0.3338 | 0.5885 |
| 30 | 0.2839 | 0.3407 | 0.5943 |
| 35 | 0.2796 | 0.3609 | 0.5945 |
| 40 | 0.2807 | 0.3620 | 0.6045 |




H04N

그림 1 문서건수증가에 따른 전체 F-Measure 변화 비교

| IPC | 유형 | PMI + COSINE | 제안방법 (필터링 전) | 제안방법 (필터링 후) |
|---|---|---|---|---|
| H04N | 유의어 | 0.3362 | 0.4505 | 0.4189 |
| | 이형어 | 0.3494 | 0.5816 | 0.6257 |
| | **전체** | **0.3428** | **0.5161** | **0.5223** |
| G06F | 유의어 | 0.3643 | 0.5079 | 0.4212 |
| | 이형어 | 0.4541 | 0.5962 | 0.6042 |
| | **전체** | **0.4092** | **0.5521** | **0.5127** |




H04N G06F

그림 2 연관단어 필터링 적용 여부에 따른 F-Measure 변화

평가 결과, 5,000 개의 특허정보만 이용하여 대체어를 추출하였을 경우, 제안방법이 기존의 대체어 추출 연구들(Context Window Overlapping[3], PMI+COSINE[1])보다 F-Measure에 있어서 최대 약 8배 정도 높은 성능을 지니는 것으로 나타났다. 향후에는 본 연구에서 이용한 특허정보 데이터를 내에 존재하는 띄어쓰기 오류 및 복합어 처리 문제 등을 처리하기 위한 연구를 수행할 필요가 있으며, 본 연구에서 제안한 연관단어필터링 기법의 적용에 있어서 유의어 유형의 Recall이 하락하는 원인을 규명하고, 이를 보완하여 최종적인 '대체어 결정 함수'로 발전시킬 필요가 있다.

## 참고문헌

[1] J. Baik and S. Kim and S. Lee, "Automatic Construction of Alternative Word Candidates to Improve Patent Information Search Quality", *Journal of KIISE:Software and Applications*, vol.36, no.10, pp.861-873, 2009.(in Korean)

[2] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL", in Proceedings of the Twelfth European Conference on Machine Learning, 2001.

[3] Ruiz-Casado, M. and Alfonseca, E. and Castells, P., "Using Context-Window Overlapping in Synonym Discovery and Ontology Extension", in Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP-2005, 2005

[4] PD Turney, "Similarity of Semantic Relations", *Computational Linguistics*, Vol.32, No.3, pp.379-416, 2006

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.