병렬말뭉치를 이용한 대체어 자동 추출 방법

(Automatic Extraction of Alternative Words using Parallel Corpus)

백 종 범 † 이 수 원 ††
(Jongbum Baik) (Soowon Lee)

요 약 정보 검색에 있어서 동일 객체를 다양한 표기로 기술하는 문제는 시스템의 성능을 저하시키는 요인이 된다. 본 연구에서는 이러한 문제를 해결하기 위하여 특허 정보의 국/영문 제목을 병렬말뭉치로 이용하여 대역어 뭉치를 추출하고, 이를 각 단어의 특징(Feature)으로 이용하여 대체어 목록을 자동 추출하는 방법을 제안한다. 또한 대체어 목록 내에 대체어가 아닌 다수의 연관단어들이 포함되는 문제점을 해결하기 위하여 국문 제목에서 추출한 연관단어 뭉치를 이용하여 대체어 목록 내 연관단어들을 필터링하는 방법을 제안한다. 평가결과에 따르면 본 연구에서 제안한 방법이 기존의 대체어 추출 방법들보다 더 우수한 것으로 나타났다.

키워드 : 동의어, 대체어, 병렬말뭉치, 텍스트 마이닝

Abstract In information retrieval, different surface forms of the same object can cause poor performance of systems. In this paper, we propose the method extracting alternative words using translation words as features of each word extracted from parallel corpus, korean/english title pair of patent information. Also, we propose an association word filtering method to remove association words from an alternative word list. Evaluation results show that the proposed method outperforms other alternative word extraction methods.

Key words : Synonym, Alternative Word, Parallel Corpus, Text Mining

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 불가능하다. 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 표 1과 같이 이형어, 대역어, 유의어로 분류한다. 특히, 본 연구에서는 세 가지 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[1-3]. 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. 예를 들어 '기능'이라

표 1 대체어 유형의 정의

유 형	정 의
이형어	기준단어와 동일한 대상을 다른 철자로 표기한 경우
대역어	영어로 표기된 기준단어에 대한 한글 표기 혹은 그 반대의 경우
유의어	기준단어와 비슷한 의미를 지닌 단어

· 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
· 이 논문은 2010 한국컴퓨터종합학술대회에서 '병렬말뭉치를 이용한 대체어 자동 추출 방법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 숭실대학교 컴퓨터학과
jbb100@ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학과 교수
swlee@ssu.ac.kr

논문접수 : 2010년 8월 12일
심사완료 : 2010년 10월 22일

는 단어는 '텔레비전'을 설명하는 연관단어인 동시에 '단말기', '수신기' 등 수많은 다른 단어들과도 동시출현빈도가 높은 연관단어들이다. 그러므로 '텔레비전'을 다른 단어와 구별 지어줄 특징으로서의 역할을 수행하기에 부족함이 있으며, 특징 선택 과정에 있어서 '기능'과 같이 여러 단어들과의 동시출현 빈도가 높은 연관단어들을 모두 제거하면 Overfitting문제가 발생하여 오히려 대체어 추출 성능이 저하되는 문제가 발생한다. 또한 학습 문서의 개수가 충분히 확보되지 못한 경우에는 연관단어 뭉치 내에서 특징을 선택하는 작업이 더욱 어려워진다. 본 연구에서는 이러한 연관단어 뭉치의 단점을 극복하기 위하여 병렬말뭉치로부터 추출한 대역어 뭉치를 각 단어의 특징으로 이용한다.

문맥 정보를 이용한 대체어 추출 방법 외에도 사전을 이용한 방법[4-6], 문장 패턴을 이용한 방법[7,8] 등이 존재하지만 이러한 방법들은 한정된 어휘 및 패턴 내에서만 대체어 추출이 가능하다는 단점이 있다. 이와 같은 방법으로는 본 연구에서 정의한 이형어를 추출할 수 없으므로 본 연구에서는 이러한 연구들을 추가적으로 고려하지 않는다.

대체어 자동 추출을 위하여 본 연구에서 제안하는 방법은 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.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 연구에서 제안하는 방법을 설명하며, 3장에서는 제안방법에 대한 실험결과를 논의한다. 마지막으로 4장에서는 결론을 기술하고 향후 연구를 제시한다.

2. 대체어 자동 추출 시스템

본 연구에서 제안하는 대체어 추출 시스템은 그림 1과 같이 총 다섯 단계로 이루어진다. 본 장의 각 절에서는 제안하는 방법을 단계별로 설명한다. 단, 특허정보수집 및 전처리 단계의 경우에는 방법적으로 특이한 점이 없으므로 본 장에서 언급하지 않는다.

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(그림 2). 본 단계에서 추출하는 대역어 뭉치는 대체어 추출 단계(2.2절)에서 이용되며, 연관단어 뭉치는 연관단어 필터링 단계(2.3절)에서 이용된다.

그림 2 대역어 뭉치와 연관단어 뭉치

본 연구에서는 대역어 뭉치를 추출하기 위해서 그림 3과 같이 국문 제목과 영문 제목 간에 동시 출현한 '한글-영어 단어 쌍'의 빈도를 이용하여 Jaccard 상관계수를 계산한다[9]. 이는 기준단어와 함께 가장 많이 출현하는 대역어를 찾기 위한 과정이다. 대역어 뭉치 추출 과정에 있어서 D_{w1}은 국문 제목 내에서 '한글 단어' w_1이 출현한 문서들의 집합으로 정의하고, D_{w2}는 영문 제목 내에 '영어 단어' w_2가 출현한 문서들의 집합으로 정의한다.

$$\text{Jaccard}(w_1, w_2) = \frac{|D_{w1} \cap D_{w2}|}{|D_{w1} \cup D_{w2}|}$$

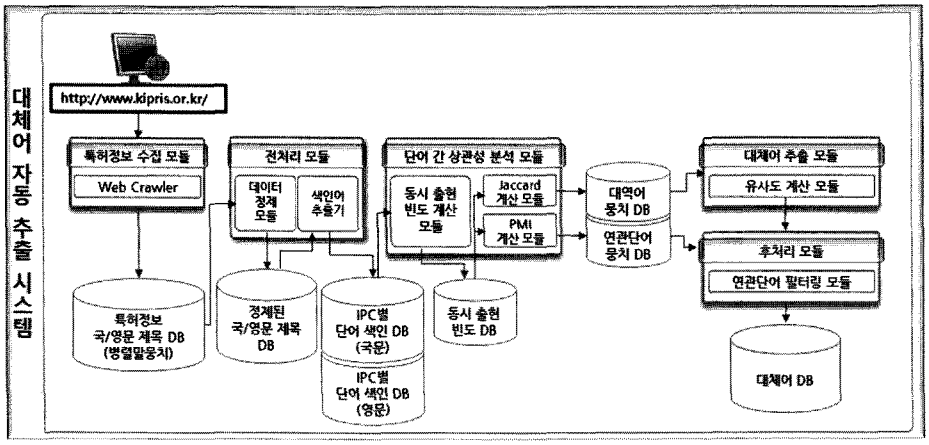

그림 1 대체어 자동 추출 시스템 구조도

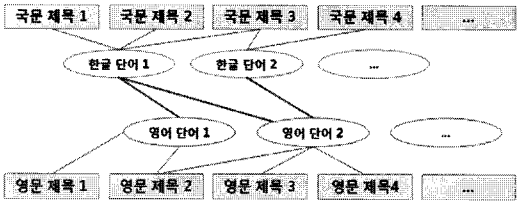

그림 3 병렬말뭉치를 이용한 대역어 뭉치 추출 과정

표 2는 Jaccard 상관계수를 계산하여 추출한 대역어 뭉치의 일부이다.

표 2 '데이타' 및 '데이터'에 대한 대역어 뭉치 (Jaccard > 0.03)

기준단어	대역어	Jaccard
데이타	DATA	0.0588
데이터	DATA	0.6260
데이터	METHOD	0.0710
...	...	...

또한 연관단어 뭉치를 추출하는 과정에 있어서는 그림 2와 같이 국문 제목 내 출현 단어 간 PMI(Point-wise Mutual Information)[1]를 계산한다. 연관단어 뭉치 추출 과정에 있어서 D_{w1}, D_{w2}는 모두 국문 제목 내에서 '한글 단어' w_1 및 w_2가 출현한 문서들의 집합으로 정의한다. 표 3은 PMI를 계산하여 추출한 연관단어 뭉치의 일부이다.

$$PMI(w_1, w_2) = \log \frac{p(D_{w1} \cap D_{w2})}{p(D_{w1})p(D_{w2})}$$

표 3 '데이터'에 대한 연관단어 뭉치(PMI > 0)

기준단어	연관단어	PMI
데이터	스레드	4.9481
데이터	사태	4.6854
데이터	캐싱	4.5856
데이터	포스	4.5329
...	...	...

2.2 대체어 추출

본 연구에서는 대체어 목록을 추출하기 위하여 기준단어와 조합 가능한 모든 단어 간의 코사인 유사도(Cosine Similarity)[4]를 계산한다. 두 단어 간의 코사인 유사도를 계산하기 위해서는 각 단어를 설명하는 특징벡터가 존재해야 한다. 본 연구에서는 3.1절에서 추출한 대역어 뭉치를 각 단어의 특징벡터로 정의하여 단어 간의 유사도를 계산한다. 표 4는 단어 간 코사인 유사도를 계산하여 추출한 대체어 목록의 예이다. 표 4에 따르면 '데이터'의 대체어인 '데이타'가 목록의 6번째 순위에 등장하는 것으로 나타났다.

표 4 '데이터'에 대한 대체어 목록

기준 단어	대체어	코사인 유사도	PMI
데이터	방법	0.5422	2.5
데이터	처리	0.5305	4.79
데이터	장치	0.5287	2.21
데이터	시스템	0.5181	2.39
데이터	방송	0.4991	4.2
데이터	데이타	0.4936	0
...	...	...	...

2.3 연관단어 필터링

코사인 유사도를 이용하여 추출한 대체어 목록 내에는 표 4에서 나타나는 바와 같이 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 3.1절에서 추출한 연관단어 뭉치를 이용한 연관단어 필터링 방법을 제안한다. 본 단계의 기본 아이디어는 [3]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 상관 관계, 0보다 작을 때에는 음의 상관 관계를 지닌 것으로 판단한다. 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. 표 5는 표 4를 필터링한 결과로서 연관단어 뭉치가 제거된 후, '데이터'의 대체어로 '데이타' 한 개만 남는 것으로 나타났다.

표 5 '데이터'의 대체어 목록에 대한 연관단어 필터링 수행 결과 (PMI ≤ 0)

기준단어	대체어	코사인유사도	PMI
데이터	데이타	0.4936	0

3. 실험 및 결과

3.1 실험 데이터

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.

3.2 평가지표

본 연구에서는 3.1절의 데이터로부터 추출한 대체어의 품질을 평가하기 위하여 2개의 IPC 분류(H04N, G06F)에 대하여 수작업으로 평가지표를 구축하였다. 구축된 평가지표는 대체어 유형별로 각각 대역어 사례 601건, 이형어 사례 52건, 유의어 사례 191건으로 구성되어 있다.

3.3 평가방법

본 연구에서는 제안방법의 성능 평가를 위한 척도로서 MAP(Mean Average Precision)[4]와 Recall을 이용하였다. 또한 MAP와 Recall을 종합적으로 고려한 성능을 평가하기 위하여 F-Measure를 계산함으로써 시스템의 최종적인 성능을 평가하였다.

$$MAP = \frac{\sum_{r=1}^{N} P(r) \times rel(r))}{\text{대체어 평가지표 내 대체어 사례의 개수}}$$

$$P(r) = \frac{\text{r 번째 혹은 더 높은 순위에서 대체어 사례가 등장한 횟수}}{r}$$

MAP에서 r은 순위(Rank)를 의미하며, N은 추출된 대체어 목록의 개수를 의미한다. 또한 rel(r)은 특정 순위에서 대체어 사례(Relevant Alternative Word)가 출현할 경우에만 활성화되는 이진함수(Binary Function)이다. 마지막으로 P(r)은 특정 순위에서의 정확율(Precision)을 의미한다.

$$Recall = \frac{\text{대체어 목록 내에 출현한 대체어 사례 개수}}{\text{대체어 평가지표 내 대체어 사례의 개수}}$$

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

3.4 평가결과

3.4.1 타 시스템과의 성능 비교

본 연구에서는 제안방법의 객관적인 성능을 평가하기 위해서 문맥정보를 이용하여 대체어를 추출하는 기존 연구들 중Context Window Overlapping(CWO[ONE_WAY])[2] 및 PMI+COSINE[3] 비교 대상으로 선정하여 비교평가를 수행하였다. 그림 4는 실험 데이터(특허정보)의 건수를 5,000건 단위로 최대 40,000건까지 증가시켜가며 각 대체어 유형별로 기존연구와 제안방법의 성능 변화를 비교한 결과이다. 평가 결과에 따르면 대역어, 유의어, 이형어 등 모든 유형에 있어서 제안방법이 더 높은 성능을 보이는 것으로 나타났다.

특히 대역어 유형에 있어서 특허정보 건수와 무관하게 약 75%정도의 F-Measure를 유지하는 것으로 나타났다.

유형 구분 없이 전체적인 F-Measure를 비교한 결과에서도 제안방법이 기존 방법들에 비하여 평균적으로 약 30%포인트 정도 향상된 것으로 나타났다(그림 4).

3.4.2 연관단어 필터링 적용 여부에 따른 성능 비교

본 실험은 본 논문의 3.3절에서 제안한 '연관단어 필터링'기법의 성능을 평가하기 위하여 3.1절에서 언급한 데이터 전체를 이용하여 수행되었다.

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

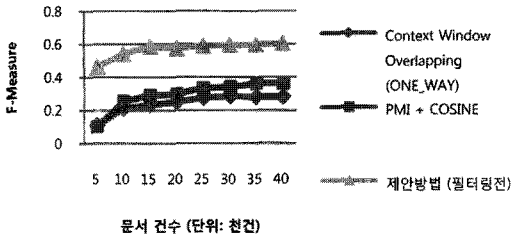

그림 4 특허정보 건수 증가에 따른 전체 F-Measure 변화 비교(H04N)

그림 5는 Precision과 Recall을 종합하여 F-Measure를 계산한 결과로서 유의어에 있어서는 약 4~8%포인트 정도 성능이 하락한 것으로 나타나며, 이형어에 있어서는 약 1~7% 포인트 정도 향상된 것으로 나타났다. 전체적으로는 H04N 분류에서는 약 1% 포인트 정도 성능이 향상된 것으로 나타났으나, G06F에 있어서는 4% 포인트 정도 성능이 하락한 것으로 나타났다.

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

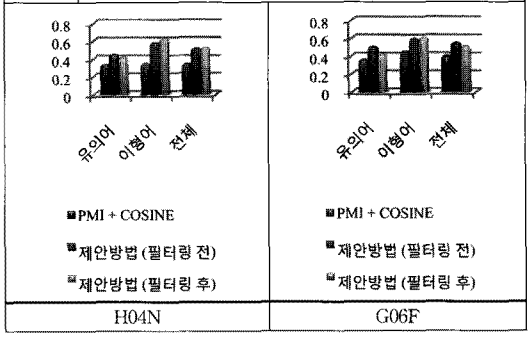

그림 5 연관단어 필터링 적용 여부에 따른 F-Measure 변화

4. 결론 및 향후 연구

본 연구에서는 기존의 문맥정보를 이용한 대체어 추출 연구들의 성능이 입력 문서뭉치의 건수에 영향을 받는 문제점을 분석 및 정의하고, 이를 극복하기 위하여 병렬말뭉치로부터 추출한 대역어 뭉치를 이용하여 대체어를 추출하는 방법을 제안하였다. 또한 최종적인 대체어 목록을 자동으로 결정하기 위한 연관단어 필터링 기법을 제안하였다.

제안방법은 병렬말뭉치로부터 Jaccard 상관계수를 이용하여 대역어 뭉치를 추출하였다는 점과 추출된 대역어 뭉치를 각 단어의 특징으로 이용하여 코사인 유사도를 계산한 후, PMI를 이용한 연관단어 필터링 기법을 적용하여 대체어 목록의 품질을 향상시켰다는 점에 있어서 기존 연구들과 차별된다.

평가 결과, 5,000 개의 특허정보만 이용하여 대체어를 추출하였을 경우, 제안방법이 기존의 대체어 추출 연구들(Context Window Overlapping, PMI+COSINE)보다 F-Measure에 있어서 최대 약 8배 정도 높은 성능을 지니는 것으로 나타났다. 또한 본 연구에서 제안한 연관단어필터링 기법을 적용한 결과, MAP에 있어서 약 10% 포인트 정도 향상된 것으로 나타났다. 그러나 유의어 유형의 Recall에 있어서는 약 13% 포인트 정도 하락한 것으로 나타났다.

본 연구에서 이용한 특허정보 데이터를 관찰한 결과, 띄어쓰기 오류 및 복합어 처리 문제 등 전처리가 필요한 문제점들이 발견되었다. 향후에는 이러한 부분이 추가적으로 연구되어야할 필요가 있다. 또한 본 연구에서 제안한 PMI를 이용한 연관단어필터링 기법의 적용에 있어서 유의어 유형의 Recall이 하락하는 원인을 규명하고, 이를 보완하여 최종적인 '대체어 결정 함수'로 발전시킬 필요가 있다.

참 고 문 헌

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

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

[3] J. Baik, 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)

[4] Pierre P. Senellart and Vincent D. Blondel, "Automatic discovery of similar words," *in Survey of Text Mining*, Springer, 2003.

[5] Jon M. Kleinberg, "Automatic construction of networks of concepts characterizing document databases," *Journal of the ACM*, vol.46, no.5, pp.604-632, 1999.

[6] Vincent D. Blondel and Pierre P. Senellart, "Automatic extraction of synonyms in a dictionary," Presented at the TextMining Workshop, Arlington, Virginia, 2002.

[7] John McCrae and Nigel Collier, "Synonym Set Extraction from the Biomedical Literature by Lexical Pattern Discovery," *BMC Bioinformatics*, vol.9, no.159, 2008.

[8] Rema Ananthanarayanan, Vijil Chenthamarakshan, Prasad M Deshpande, and Raghuram Krishnapuram, "Rule based Synonyms for Entity Extraction from Noisy Text," *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data In AND '08*, pp.31-38, New York, NY, USA, 2008.

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