형태정보를 이용한 대역어 군집화 및 적합대역어 선정

구회관°* 정한민** 이미경** 성원경**
°*과학기술연합대학원대학교 응용정보과학**한국과학기술정보연구원 차세대정보시스템연구실
{hkkoo°, jhm, jerryis, wksung}@kisti.re.kr

Translation Clustering and Adequate Translation Selection by Surface Form

Heekwan Koo°* Hanmin Jung** Mikyoung Lee** Won-Kyung Sung**
°*Practical Information Science, UST
**Information System Research Lab., KISTI

요 약

본 논문은 자동적인 언어기반자원구축을 위해 신문 말뭉치에서 괄호를 이용하여 추출한 대역어쌍들을 군집화하고 각 군집에서 적합대역어를 선정하는 방법을 제안한다. 기존 연구에서 주로 제시된 음차표기어 대역쌍 추출 방법은 완전한 형태의 영어원어 자소 정보를 이용하기 때문에 약어는 고려대상에서 제외되었다. 그러나 약어형태의 영어원어가 신문에서는 약 82%를 차지하기 때문에 이를 처리할 방법이 필요하다. 따라서 본 논문에서는 바이그램을 기본으로 하는 형태정보를 이용하여 적합대역어를 선정하고 이와 형태정보를 공유하는 한국어대역어쌍들을 군집화한다. 또한, 음차표기어와 두문자어에 대한 처리를 추가하여 적용범위를 넓힌다. 실험을 위하여 신문말뭉치에서 추출한 대역어쌍 1,806개 중 영어원어를 기준으로 한국어대역어의 수가 5개 이상인 대역어쌍 집합 200개를 선정하였다. 본 논문에서 제시한 방법으로 측정한 결과, 대역어 군집화에 대해서는 74%의 정확율과 65%의 재현율을, 적합대역어 선정에 대해서는 97%의 정확율을 보였다.

1. 서론

과학기술의 발전으로 용어는 지속적으로 변화와 발전 과정을 거친다. 이런 이유로 용어의 현 상태를 적절하게 반영할 수 있는 언어자원구축 방법에 대한 요구가 증가하고 있다. 신문말뭉치를 기반자원으로 활용하는 것도 유용한 방법들 중 하나이다. 대역사전을 포함한 지식베이스를 구축하기 위해서 신문말뭉치로부터 괄호 등을 포함하는 대역어쌍들을 추출할 수 있다.

신문말뭉치에서는 "인터넷서비스업체(ISP)," "컨텐츠(CONTENTS)"와 같은 형태의 한국어-영어 대역어쌍을 흔히 볼 수 있다. 전자신문에서 수집한 전체 대역어쌍들을 분석해 보면, "ISP"와 같이 영어약어를 사용하는 비율이 전체비중에 82%를 차지하며, 나머지 18% 정도는 "CONTENTS"와 같은 완전한 형태의 영어원어를 사용한다. 또한, 자동 추출된 한국어대역어 수가 5개 이상인 대역어쌍 집합을 대상으로 하여 분석을 하면 영어약어 비중이 98%까지 높아진다. 이러한 결과는 영어약어가 한국어 대역어의 애매성을 증가시킬 수 있다는 사실을 의미하며, 이들에 대한 처리방안이 필요하다는 사실을 지적한다.

자동 음차 표기는 크게 영어 알파벳을 한글 자소로 변환하는 직접 방식과, 영어 알파벳을 발음 사전을 이용하여 음소로 바꾸고 다시 한글 자소로 변환하는 피봇 방식이 있다 [2]. 대역쌍 추출은 대부분 직접방식으로 구현하며, 구현 방법에 의해 확률 기반 모델, 결정 트리 기반 모델, 음차 표기 네트워크 기반 모델 등으로 나뉠 수 있다 [5].

기존 연구들은 "컨텐츠(CONTENTS)"와 같은 완전한 형태의 영어원어 대역어 쌍들을 대상으로 대역어와 음차표기 대역쌍 추출에 집중하였다[1] [4] [3] [5] [2]. 대역어 추출에는 외부정보가 없으면 추출하기 어렵기 때문에 주로 사전을 이용한다 [5]. 사전을 이용할 경우, 미등록어가 사전에 의해 번역이 되지 않는 문제를 해결하기 위해 음운유사도로 음차표기어를 추출하고 대역어의 부분일치도를 이용하였다. 이렇게 사전을 이용하려면 완전한 형태의 원어가 필요하지만 영어원어가 약어의 형태로 구성된 대역쌍에는 적용하기 힘들다.

영어원어가 약어이면 기존의 사전이나 부분대역어 일치 방법들 대신에 서로간의 형태적 유사성을 이용하여 군집화하는 방안을 고려할 수 밖에 없다. 실제로 신문 말뭉치에서 추출된 대역어쌍들을 살펴보면, "EC"라는 영어원어에 대해 "전자상거래," "유럽위원회," "전해콘덴서" 등을 중심으로 형태적 군집이 형성되는 것을 알 수 있다. 본 논문은 이러한 형태정보를 이용하여 적합대역어들을 선정하고, 이들을 중심으로 대역어들을 군집화하는 방안을 제안한다. 바이그램을 기본으로 하여 현재 남아 있는 대역어집합에서 적합대역어 하나를 선정하고 이와 형태적 공유를 가지는 대역어들을 이 적합대역어의 군집에 포함시킨다. 이러한 방식을 되풀이하여 군집화를 완성해간다. 또한, 음차표기어와 두문자어에 대한 처리를 위해 적합대역어와 후보대역어들을 스캐닝하는 방식으로 비교한다.

2. 적합대역어 선정 및 대역어 군집화

```
ExtractTranslationCluster(S) {
        C1 = FindTransliteration(S);
        S = S - C1;
        for( k = 2; S != NULL; k++)
        {
                termk = FindFittest(S);
                Ck = FindCluster(S, termk);
                Ck = Ck + FindAcronym(S, termk);
                S = S - Ck;
        }

        return C;
}
```

그림 1. 적합대역어 선정 및 대역어 군집화 알고리즘

그림 1은 적합대역어를 선정하고 이를 중심으로 군집화하는 알고리즘을 보여준다. 집합 S는 하나의 영어원어가 가지는 한국어대역어들이다. 이들은 말뭉치로부터 괄호와 같이 출현하는 대역어들을 자동 추출하여 수집된다. FindTransliteration()은 한국어대역어들로부터 음차표기어들을 찾아 하나의 별도 군집 (C_1)을 만든다 (2.1 참조). 음차표기어들을 별도로 군집화하는 이유는 음차표기어 자체는 대역정보를 가지지 못하여 의미적으로 군집화될 수 있는 성격이 아니기 때문이다. 번역된 대역어들은 번역 과정에서 의미를 가질 수 있지만, 음차표기어들은 단순히 영어원어를 발음나는 대로 옮겨놓은 것에 불과하기 때문에 대역 애매성이 해소되지 않은 상태이다.

음차표기어들을 집합 S에서 제거한 후, 집합 내의 모든 대역어들이 군집화가 될 때까지 반복적으로 군집화하는 과정을 수행한다. 첫번째 단계로 현재의 집합 내에서 적합대역어를 찾는다 (FindFittest()). 적합대역어 ($term_k$)를 찾기 위한 기본 가정은 "적합대역어는 대역어 집합 내에서 형태적으로 가장 많이 공유된다."와, 공유 정도가 비슷한 경우에는 "길이가 짧을수록 잘못 추출된 부분을 포함하지 않는다."는 것이다. 식 (1)은 대역어 집합 내에서 각 대역어에 대해 형태적 공유 정도를 측정하는 식으로 이 값이 가장 큰 대역어를 현재 집합 내에서의 적합대역어로 선정한다.

$$T_i = \frac{\sum_{j=1}^{n} BI_{ij}}{|X_i|} \quad \text{where } i \neq j \qquad (1)$$

n개의 대역어들을 갖는 집합에서 대역어 X_i의 형태적 공유 정도 (T_i)는 자신을 제외한 나머지 대역어들과 공유하는 바이그램 수 (BI_{ij})의 합을 자신의 길이 ($|X_i|$)로 나눈 값이다. 예를 들어, "PI"라는 영어원어가 "프로세스혁신," "업무프로세스혁신," "경영혁신"을 대역어로 가진다면, "프로세스혁신"은 "업무프로세스혁신"과 5개의 바이그램을 공유하고, "경영혁신"과 1개의 바이그램을 공유하므로, 형태적 공유 정도 값이 6/5이 된다. "업무프로세스혁신"은 6/7이 되며, "경영혁신"은 1/3이 된다. 결국, "프로세스혁신"이 현재 대역어 집합 내에서 적합대역어로 선정된다.

FindCluster()는 선정된 적합대역어와 형태적 공유를 가지는 대역어들을 발견하여 하나의 군집으로 만든다. 상기 예에서 나머지 두 대역어 모두 "프로세스혁신"과 형태적 공유를 가지므로 하나의 군집이 된다. FindAcronym()은 적합대역어의 두문자어가 존재하는지를 검사하여 해당되는 경우에 군집에 추가한다 (2.2 참조) 이러한 과정을 반복하여 대역어 집합 S가 NULL이 되면 군집화를 중단하고, 현재까지의 군집들 ($C = \{C_1, C_2, \ldots\}$)을 출력한다.

2.1 음차표기어 군집화

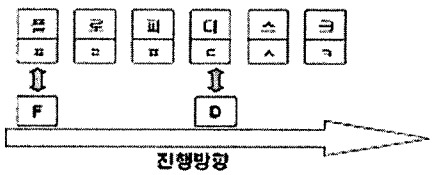

그림 2. 원어와 음차표기어와의 매칭 예

음차표기어를 추정하기 위해 KODEX [1]에서 제안한 방식과 유사하게 초성을 직접적으로 변환하는 방식을 사용하였다. 집합 S내의 모든 대역어들을 음절로 분할하고, 각 음절의 초성을 추출하여 영어원어의 진행방향으로 초성들과 대응하는지 확인하는 방법을 이용해 음차표기어를 발견한다. 그림 2는 이 과정을 도식화하여 설명한다. "FD"와 "플로피디스크"를 예로 든다면, "플로피디스크"는 "ㅍㄹㅍㄷㅅㅋ" 형태의 초성열로, "FD"는 "ㅍㄷ"이라는 초성열로 변환된다. 이들을 좌에서 우로 진행하면서 매칭하여 일치하면 한국어대역어를 음차표기어로 인정한다. "FD"와 같이 자음만으로 구성된 영어원어는 KODEX에서 제기된 방법으로 추정이 가능하지만, 영어약어는 일반적으로 길이가 짧고 그 중 모음(들)이 포함되어 있을 수 있기 때문에 (예. "CE(시스템엔지니어)") 모음들에 대해서도 대응 가능한 한글 모음을 고려해야 한다.

2.2 두문자어 추정

두문자어 추정은 적합대역어 선정과 군집화가 이루어진 후에 대역어집합 S에 대해 적합대역어와 같은 음절 순서를 가지는 용어들을 찾는 과정으로 이루어진다. 그림 3은 "한국과학기술원"이 적합대역어일 때 "과기원"이 두문자어임을 추정하는 과정을 보여준다.

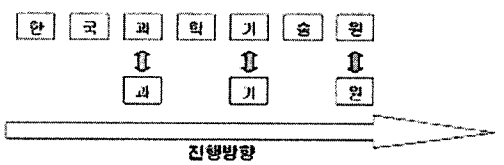

그림 3. 두문자어 추정 예

3. 실험

전자신문 말뭉치로부터 한국어와 영어가 괄호로 연결되어 있는 대역어쌍들을 자동으로 추출한 후, 영어원어를 기준으로 재정렬하여 1,806개의 영어원어에 대한 대역어집합들을 생성하였다. 이들 중 집합의 크기가 5개 이상인 200개의 대역어집합들을 본 실험에서 사용하였다 (전체 한국어대역어 수는 2,253개). 정답군집 생성을 위해 수작업으로 대역어들에 태깅을 하였다. 정답 군집을 태깅할 때, 형태적 유사성뿐만 아니라 의미적 유사성을 모두 반영하였다. 즉, 형태적으로 공유하는 부분이 없더라도 같은 의미를 가지는 대역어군인 경우 동일한 군집으로 인정하였다. 대역어 군집에서의 적합대역어는 하나 이상 존재할 수 있다. 예를 들어, "NFS"는 "미과학재단," "미국과학재단," "전미과학재단"을 동시에 적합대역어로 가질 수 있으며, 실험에서는 이들 중 하나를 적합대역어로 찾으면 정답으로 인정하였다.

실험 결과는 표 1과 같다. 성능 비교 평가를 위해 Dice 계수 및 Jaccard 계수를 이용하였다. 적합대역어 정확율은 적합대역어로 선정된 용어가 올바른 지를

보여주는 값이다. 각 알고리즘이 추천한 군집들이 가지는 적합대역어들의 적합성을 결정한 것으로, 올바른 적합대역어를 전체 적합대역어 수로 나눈 값이다. 군집화 재현율은 정답 군집들 중에서 각 알고리즘이 제시한 군집들과 정확히 일치하는 군집들의 비율이며, 군집화 정확율은 각 알고리즘이 제시한 군집들 중에서 올바른 군집들의 비율이다. 본 알고리즘은 비교 대상 알고리즘들에 비해 군집화에 대해서 약간 좋은 성능을 보인다. 특히, 적합대역어 정확율에 있어서는 비교 대상 알고리즘들보다 13% 높은 성능을 보여준다.

표 1. 본 알고리즘과 Dice 계수[1], Jaccard 계수[2]와의 비교 결과

	Dice 계수	Jaccard 계수	본 알고리즘
군집 수	621개	621개	615
평균 군집 크기	3.628	3.628	3.663
적합대역어 정확율	84.219% (523/621)	84.219% (523/621)	97.236% (598/615)
군집화 재현율	62.267% (434/697)	62.267% (434/697)	65.136% (454/697)
군집화 정확율	69.887% (434/621)	69.887% (434/621)	73.821% (454/615)
군집화 F-measure	65.857%	65.857%	69.207%

본 알고리즘이 비교 대상 알고리즘들보다 적합대역어 정확율에 있어서 높은 성능을 보이는 이유는 실험 대상이 자동 추출 방식에 의해 구성되어 추출 오류들을, 특히 적합대역어에 군더더기가 붙은 형태들을 다수 포함하고 있어서 이들을 배제할 수 있도록 길이 정보를 차별적으로 적용하였기 때문인 것으로 보인다. 적합대역어 선정 오류 중 대표적인 예를 보면 다음과 같다. "EMC"라는 영어원어와 "반도체봉지재," "반도체봉지제," "반도체용봉지제," "봉지재"를 포함하는 대역어군집에서 본 알고리즘에 의해 적합대역어를 선정하면, "봉지재"가 선정이 되는 문제점을 보인다. 이러한 현상은 본 알고리즘이 짧은 길이의 대역어를 선호하기 때문이며, 이를 해결하기 위해 적절한 대역어 길이에 대한 보정을 알고리즘에 반영해야 할 것이다.

군집화의 성능을 저하시키는 몇몇 대표적인 경우들을 살펴보면 다음과 같다. "WTO"는 "세계무역기구"와 "세계관광기구"라는 다른 의미의 대역어를 가지지만, "세계"와 "기구"라고 하는 형태적 유사성을 가진다. 또한 "TRS"의 대역어의 정답을 "주파수공용통신" 과 "국가기관통합무선망"으로 나눌 때, 두 군집간에 형태적인 정보가 같게 나타나는 "국가재난통신망"이라는 용어는 처음 주파수공용통신이 군집화 될 때 그 군집 속에 포함되는 문제점을 가지게 된다. "시스템,기술" 등이 포함된 대역어가 추천대역어로 선정되어 군집화가 진행되면, 다른 군집들에 속하는 대역어가 그 군집에 포함되는 경우가 빈번하게 일어난다. 이러한 문제점은 향후 형태소 분석을 이용하여 정확율을 저하시키는 단어들을 불용어 처리할 수 있는 방안을 마련하는 방향으로 개선할 필요가 있다.

5. 결론

본 논문은 바이그램에 기반한 형태정보를 이용하여 적합대역어들을 선정하고 이들을 중심으로 대역어들을 군집화하는 방식을 제안함으로써 실제 신문말뭉치에 대해 적용이 가능하도록 하였다. 또한, 음차표기어와 두문자어도 적용 범위에 포함시켜 다양한 대역 현상들을 처리할 수 있도록 함으로써 대역사전을 포함한 다양한 지식베이스 구축을 위한 기본 자료로서의 가치를 높이고 있다. 구축자들은 본 시스템이 추천한 적합대역어와 대역어군집을 검토하는 것으로 업무범위를 줄일 수 있게 된다.

참고문헌

[1] 강병주, 한국어 정보검색에서 외래어와 영어로 인한 단어불일치문제의 해결, 박사학위논문, 한국과학기술원, 2001.

[2] 오종훈, 배선미, 최기선, 자동 음차표기를 이용한 영-한음차표기 대역쌍의 자동추출, *한국정보과학회 춘계학술대회*, 2004.

[3] 오종훈, 최기선, *자소 및 음소 정보를 이용한 영어-한국어 음차표기 모델*, 한국정보과학회 논문지 32(4), 2005.

[4] 이재성, 다국어 정보검색을 위한 영한 음차표기 및 복원 모델, 박사학위논문, 한국과학기술원, 1999.

[5] 이재성, 서영훈, *한영 혼용문에서 괄호 안 대역어구의 자동인식*, 한국정보처리학회 논문지 B 9-B(4), 2002.

[1] $\frac{2(X \cap Y)}{X+Y}$의 두 대역어에 대한 계산을 확장하여, 대역어 X와 모든 나머지 대역어들 간의 Dice 계수를 구하여 이들의 합으로 대역어 X의 형태적 공유 정도를 결정하였다.

[2] $\frac{X \cap Y}{X \cup Y}$의 두 대역어에 대한 계산을 확장하여, 대역어 X와 모든 나머지 대역어들 간의 Jaccard 계수를 구하여 이들의 합으로 대역어 X의 형태적 공유 정도를 결정하였다.