

음절 빈도를 이용한 외래어 명사의 인식

강승식, 전영진

국민대학교 컴퓨터학부, 첨단정보기술연구센터

e-mail: {sskang, terrius7}@cs.kookmin.ac.kr

Recognition of Foreign Nouns using Syllable Frequency

Seung-Shik Kang, Young-Jin Chun

School of Computer Science, Kookmin University and AITrc

요약

정보 검색에서 우리가 자주 쓰는 단어는 표준어인 경우도 있고 통신어인 경우도 있고 외국어인 경우도 있다. 그러나 표준어가 아닌 다른 언어로 검색을 하면 다른 결과가 나타날 수 있다. 예를 들어 컴퓨터에 관한 정보를 찾을 때 ‘컴퓨터’로 검색을 하면 다른 검색 결과가 나오게 된다. 우리나라에서 현재 쓰이고 있는 말들은 이런 애매한 발음의 외래어가 많이 생성되고 소멸된다. 그러므로 이런 외래어들을 전부 사전에 등록할 수는 없고 설사 등록한다 하더라도 용량과 검색시간만 늘어나게 된다. 본 논문에서는 검색엔진에서 이런 외래어에 대한 인식 성능을 높이기 위해 외래어 사전 없이 외래어를 인식하는 방안을 제시한다.

1. 서론

우리는 검색을 할 때 보통 한글로 검색을 한다. 그러나 우리가 쓰는 말에는 순수 한글보다는 외래어가 더 많은 것이 사실이다. 중국에서 들어온 말도 많고 서양 특히 미국에서 온 말들도 많다. 외래어란 ‘라디오’, ‘컴퓨터’, ‘라이터’처럼 외국에서 들어온 말임에도 불구하고 우리나라 말처럼 사용되는 언어를 말한다. 물론 이런 외래어에도 ‘남포’나 ‘담배’처럼 외국에서 들어온 지 너무 오래되어서 고유어처럼 인식되는 말이 있는 반면에 ‘인터넷’, ‘컴퓨터’처럼 들어온 지 얼마 되지 않아 아직 외국어의 느낌이 강한 말까지 여러 단계가 있다. 외래어란 어원적으로 외국어에서 온 말이기 때문에 넓은 의미의 외래어는 한자어도 포함된다. 그러나 좁은 의미의 외래어는 연중들의 의식 속에 외국어에서 온 말이라는 느낌이 뚜렷한, 주로 서양의 언어에서 들어온 외래어이며 한자는 배제된다. 한자는 들어온 지 오래되었고 일상생활에서 빼놓으면 말이 통하지 않을 정도이기 때문에 외국어에서 온 느낌이 별로 없으며 또한 어형이 흔들림 없이 고정되어 있다. 이에 반해 서양 언어에서 들어온 외래어는 어형이 매우 불안정한 특징이 있다. ‘텔레비전’만 하더라도 표준형인 ‘텔레비전’ 외에 ‘텔레비전’, ‘텔레비죤’ 등이 사용되는 것을 발견할 수 있고 ‘가스’는 ‘개스’라고 쓰는 사람도 있다. 이런 말들이 너무 흔하고 편하게 쓰이기 때문에 사용자는 쓰면서도 이 단어가 외국어인지 외래어인지 한글인지 모르거나 구분하기 어려운 경우도 많다. 길거리 매장의 간판이나 여러 잡지들 심지어는 기사를 정확하게 전달해야 하는 신문에서 조차도 서로 다르게 표기하는 경우가 있다. 물론 외래어도 일종의 우리말이면서 외국에서 온 말이기 때문에 어떻게 발음하던 잘못된 것이라고 할 수는 없지만 어느 하나로 통일되어 있지 않기 때문에 혼란이 있을 수 있다.

물론 교육부에서 나온 외래어 표기 유�례와 영어 철자에 따른 발음법이 있긴 하지만 널리 알려져 있지 않고 정보 검색시나 일상 생활에서 사용하기 위해 일일이 찾아보지는 않는다.

사용자가 검색을 할 때 표준어가 아닌 외래어로 하게 되면 원하는 정보를 정확하게 찾을 수 없거나 다른 결과를 보여주는 경우가 많다. 예를 들어 “컴퓨터”를 “컴퓨터”로 검색을 하게 되면 서로 다른 정보가 검색이 된다. 그렇다고 사용자에게 일일이 정확한 외래어를 쓰라고 강요할 수는 없으므로 검색 엔진에서 자동으로 외래어를 인식해서 사용자가 원하는 결과를 검색해서 보여준다면 매우 효율적 일 것이다. 즉 검색을 할 때 사용자가 잘못된 외래어를 입력해도 검색엔진에서 이런 경우에 알맞은 외래어로 인식해서 사용자가 원하는 정확한 검색 결과를 보여준다면 사용자도 검색을 반복하거나 잘못된 정보를 보는 시간을 줄이고 좀더 정확한 정보를 얻을 수 있다.

그리기 위해서는 우선 검색에 쓰인 단어가 외래어인지 아닌지 판단해야하는 작업이 선행되어야 한다. 검색하는 단어는 외래어일 수도 있고 아닐 수도 있다. 이 외래어를 판단하는 것은 간단히 외래어들을 외래어 사전에 등록해서 사전 검색으로 외래어를 판단할 수도 있지만 이 외래어라는 말은 필요에 따라 생기기도 하고 없어지기도 하므로 비효율적이다. 본 논문에서는 따로 사전의 구축 없이 외래어들의 음절 출현 빈도를 조사한 표를 구성하여 이를 바탕으로 외래어 명사를 인식하는 모델을 제안한다.

2. 제안 모델

검색엔진 네이버의 “외래어 표기 유�례”에서 얻은 3,200여 개의 외래어와 여러 웹사이트 등에서 수작업으로 추출한 약 1,000여 개의 외래어들의 자모 빈도와 음절 빈도에 대해서 분석해 보고 외래어에서 쓰이는 모든 음절의 출현

1) 본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

빈도를 바탕으로 외래어 표를 구성하였다. 이에 대한 실험으로는 웹문서에서 쓰이는 단어들을 대상으로 외래어 판정을 하였다.

본 논문에서는 크게 외래어의 자모 빈도와 음절 빈도를 분석한 것과 그것들을 바탕으로 구성한 표와 마지막으로 그 표를 가지고 외래어를 판단하는 것으로 구성하였다.

2.1 외래어에 대한 자모 빈도와 음절 빈도 조사

여러 검색엔진과 여러 사이트에서 조사한 외래어들 중에서 주변에서 흔히 쓰이는 말이지만 이 말이 정확한지 부정확한지 알기 어려운 외래어들을 대상으로 자모 빈도와 음절 빈도를 조사, 분석하였고 분석한 것들을 대상으로 나름대로의 규칙을 만들어 보았다. 자모 빈도를 조사한 것은 수작업으로 외래어를 판단할 때 사용하였고 음절 빈도를 조사한 것은 외래어를 판단하기 위한 외래어 표를 구성하는데 사용하였다.

먼저, 자모 빈도 조사 결과를 보면 초성에서는 ‘ㄹ’, ‘ㅇ’, ‘ㅅ’, ‘ㅌ’, ‘ㅍ’, ‘ㅋ’ 등이 많이 나타났으며 중성에서는 ‘ㅓ’, ‘ㅡ’, ‘ㅏ’, ‘ㅗ’, ‘ㅔ’ 등이 많이 쓰였고 종성에서는 '@(없음), ‘ㄹ’, ‘ㄴ’ 등이 많이 사용되었다.

조사한 외래어들을 대상으로 초성, 중성, 종성으로 나누어서 일정한 규칙이 있는지 분석해 보고 몇 가지 나름대로 3가지의 규칙들로 나누어 보았다. 규칙에 대해 크게 분류해보면 모음에서의 초성 변화와 자음에서의 중성 변화, 모음에서의 종성 탈락과 같은 규칙들로 분류할 수 있었다.

표 1. 외래어 표기 방식

초성	중성	종성
ㅂ → ㅂ	ㅓ → ㅓ	ㄹ ↔ @
bar: 바 - 다	ㅐ ↔ ㅏ	Angora: 앙고라 - 앙골라
ㅋ → ㅋ	ㅗ → ㅗ	blouse: 블라우스-브라우스
cafe: 카페 - 깨페	ㅜ ↔ ㅜ	cleaning: 클리닝 - 크리닝
ㅅ → ㅆ	ㅓ ↔ ㅓ	blowing: 블로잉 - 브로잉
circle: 서클 - 써클	ㅏ ↔ ㅏ	gondola: 곤돌라 - 곤도라
ㅍ → ㅎ	ㅓ ↔ ㅓ	plaza: 플라자 - 프라자
file: 파일 - 파일	ㅓ ↔ ㅓ	Rolex: 롤렉스 - 로렉스
ㄱ → ㄱ	ㅓ ↔ ㅓ	clover: 클로버 - 크로버
gas: 가스 - 가스	ㅓ ↔ ㅓ	
ㅈ → ㅈ	ㅓ ↔ ㅓ	
jazz: 재즈 - 쟈즈	ㅓ ↔ ㅓ	
ㅊ → ㅊ	ㅓ ↔ ㅓ	
brooch: 브로치 - 브로찌	ㅓ ↔ ㅓ	
ㄷ → ㄷ	ㅓ ↔ ㅓ	
dubbing: 더빙 - 띠빙	ㅓ ↔ ㅓ	
ㄱ → ㅋ	ㅓ ↔ ㅓ	
catholic: 가톨릭 - 카톨릭	ㅓ ↔ ㅓ	
ㅌ → ㅌ	ㅓ ↔ ㅓ	
mantean: 맨토 - 망토	ㅓ ↔ ㅓ	
ㅊ → ㅈ	ㅓ ↔ ㅓ	
megahertz: 메가헤르츠-메가헤르즈	ㅓ ↔ ㅓ	
ㄷ → ㅈ	ㅓ ↔ ㅓ	
module: 모듈 - 모줄	ㅓ ↔ ㅓ	
	ㅓ ↔ ㅓ	
총 14가지	총 24가지	총 4가지

2.1.1 초성 표기 방식

외래어에서의 초성의 표기 방식은 초성들이 다른 형태로 표기되는 경우이다. 일반적으로 격음에서 많이 나타났으며 주로 격음에서 쌍자음으로 변화하였다. 그 예는 아

래와 같다.

ㅂ ↔ ㅃ	ex) bar : '바' - '빠'
ㅋ ↔ ㅋ	ex) cafe : '카'페 - '까'페
ㅅ ↔ ㅆ	ex) circle : '서'클 - '씨'클
ㅍ ↔ ㅎ	ex) file : '파'일 - '파'일
ㄱ ↔ ㄱ	ex) gas : '가'스 - '까'스

2.1.2 중성 표기 방식

외래어 명사 표기에서 중성의 표기법은 외래어 단어의 발음을 한글로 표기하는 방식과 관련된 경우가 많았고, 그 예는 아래와 같다.

ㅓ ↔ ㅓ	ex) 데이'터' - 데이'타'
ㅐ ↔ ㅏ	ex) '텔'런트 - '탈'랜트
ㅓ ↔ ㅓ	ex) 에어'컨' - 에어'콘'
ㅔ ↔ ㅔ	ex) 애니'메'이션 - 애니'메'이션
ㅓ ↔ ㅓ	ex) 초콜'렛' - 초콜'릿'

2.1.3 연속된 자음에서 종성 탈락

종성 탈락은 연속된 자음에서 종성들이 축약하여 탈락하는 경우이다.

ㄹ ↔ @	ex) Angora : 앙'꼴'라 - 앙'고'라
ㄴ ↔ @	ex) linen : '린'넨 - '리'넨
ㅅ ↔ @	ex) bridge : 브릿'지 - 브'리'지
ㄱ ↔ @	ex) blocking : 블'록'킹 - 블'로'킹

2.2 외래어 판단 기준을 위한 외래어의 음절 빈도 조사
외래어에서 자주 쓰이는 음절을 알면 어떤 단어가 외래어인지 판단하는데 도움이 될 것이라고 가정하고, 외래어들의 자모 빈도와 음절 빈도를 분석하여 그 결과를 바탕으로 외래어의 음절 빈도를 계산하였다. 음절 빈도를 조사한 외래어 자료는 교육부에서 1994년 1월에 발행한 '편수자료 II-1 외래어 표기 용례(일반 외래어)'에 수록된 3,207개의 웹에서 수집한 외래어 1,000여 개이다. 교육부의 외래어 자료는 주로 일반 용어를 수록하지만, 고유 명사(종족이나 언어의 이름 따위)도 수록하고 있다.

여기서 추출한 외래어의 음절을 1byte씩 읽어서 한 음절 단위로 추출한 음절들의 빈도를 외래어 표에 저장하였다. 외래어에서 많이 쓰이는 음절들이 외래어를 구성한다고 가정을 하였고 이렇게 구성한 표를 바탕으로 외래어를 판단하는데 사용하였다.

2.3 실험 문서 집합과 외래어 판단

실험 대상 문서는 본 실험실에서 무작위로 조사한 웹문서를 대상으로 하였다. 일반적으로 쓰이는 단어들을 대상으로 판단하기 위해서 여러 분야에 대한 웹문서를 선택하였고 각 문서들은 다양한 분야의 문서로 20개의 웹문서에서 20,000여 어절들을 대상으로 실험 문서를 구성하였다. 각 문서는 총 1,000개 어절로 제한하였고 외래어 비율이 서로 다른 문서들을 선택하였는데 외래어 비율은 총 3가지 군으로 나누어서 실험하였다. 먼저 0~199개의 저빈도 외래어 비율을 가지고 있는 문서와 201~399개의 평균적인 외래어 비율을 가지고 있는 문서, 400개 이상의 고빈도를 가지고 있는 문서들로 실험 문서를 나누어 보았다.

각 문서에서의 외래어 비율은 수작업으로 판단하여 외래어를 분류하였다. 외래어들에 대한 판단은 순우리말이나 한자가 아닌 단어가 섞여 있으면 외래어로 간주하였는데 외래어가 아닌 외국어를 제외하였고, 영어에서 유래된 외래어들로만 판단을 하고 그 이외 언어의 외래어는 고려

하지 않았다. 수작업으로 판단한 외래어들은 “외래어 표기 용례”에 나오는 외래어와 ‘에스콰이어’, ‘마리끌레르’, ‘코카콜라’와 같은 영어를 사용한 회사 이름이나 외국 사람의 이름, 외국 지명 등을 외래어로 분류하였고 ‘신한카드’와 같은 상호처럼 한글과 영어가 혼합되어 있는 우리나라 상호명 중에서 영어로 표기할 때 한글을 그대로 영어로 쓰는 상호들도 외래어로 분류하였다. 그리고 ‘데이터’, ‘데이터’와 같이 같은 단어를 지칭하지만 부르는 방식이 조금씩 다른 것들도 어느 하나를 버리지 않고 둘 다 외래어로 분류하였다.

외래어의 판단은 웹문서에서 형태소 분석기를 통하여 추출한 단어들(실험 문서)을 앞에서 만들어놓은 외래어 표와 한 음절씩 비교하여 모든 음절이 외래어 표에 있는 음절과 일치하면 그 단어는 외래어라고 판단하였다.

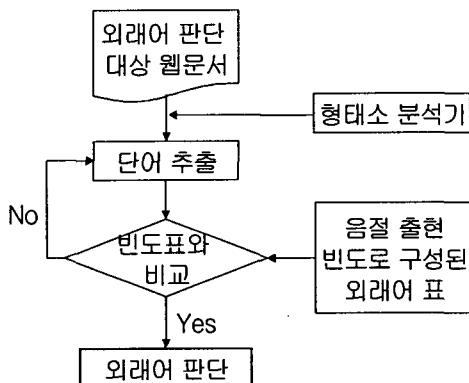


그림 1. 외래어 판단 모델 구성도

3. 실험 및 평가

정보 검색에서 일반적으로 사용되는 단어들을 대상으로 실험하기 위해 여러 분야의 웹문서를 무작위로 선택하여 그 웹문서에서 추출한 단어들로 평가를 하였다. 웹문서는 총 20개 문서에서 20,000 어절이 사용되었고 각 문서는 총 1,000개 어절로 따로 수작업을 통하여 구성하였다. 이 문서를 국민대학교 형태소 분석기를 이용하여 단어들을 추출하였고 중복된 단어들은 제거하였다. 이렇게 추출한 단어가 외래어인지 판단하기 위해 한 음절씩 읽어서 외래어 음절인지를 검사한다. 모든 음절들이 외래어 음절로 구성된 단어이면 그 단어는 외래어로 1차 판정을 하였다. 그러나 단순히 이렇게만 한다면 성능이 좋지 않기 때문에 몇 가지 제약 조건을 주어서 실험하였다. 먼저 외래어로 판정된 단어의 각 음절의 빈도수에 대하여 최저 음절 빈도수와 최대 음절 빈도수, 그리고 그 단어의 평균 음절 빈도수를 구한다. 1차 판정에서 외래어로 판단된 단어에 대해 그 단어의 각 음절이 최저 음절 빈도수와 평균 음절 빈도수를 적용하였다. 즉, 먼저 각 음절에 대한 최저음절 빈도수로 판단을 하는 실험을 해보고 여기에 다시 그 단어의 평균 빈도수를 추가로 이용하여 판단을 하는 실험을 실시하였다.

예를 들어, ‘테이블’이라는 단어에 대해서 보면 ‘테’라는 음절은 외래어 표에서 10번의 빈도수를 가지고 있고 ‘이’ 음절은 20번, ‘블’ 음절은 2번의 빈도수를 가지고 있다고 가정하면, 여기서 최저 음절 빈도수는 ‘블’ 음절의 3이고 최대 음절 빈도수는 ‘이’ 음절의 20이다. 또 평균 음절 빈도수는 $(10+20+3)/3$ 으로 11이 된다. 이렇게 구한 빈도

수들을 바탕으로 최저 음절 빈도수와 최대 음절 빈도수, 평균 음절 빈도수의 조건을 주고 성능 평가를 하였다. 각 조건은 최저 음절 빈도수가 2, 3, 5, 7, 10 일 때 평균 음절 빈도수를 2, 5, 7, 10, 13 으로 실험하였다. 즉, 한 문서 당 25개의 조건을 부여해서 25번의 실험을 하였다. 성능 평가를 위해서 정보 검색에서 많이 쓰이는 정확률과 재현율을 사용하였다.

$$\text{정확률} = \frac{\text{시스템이 정확하게 판단한 외래어 수}}{\text{총 외래어 어절 수}} \times 100$$

$$\text{재현율} = \frac{\text{시스템이 정확하게 판단한 외래어 수}}{\text{시스템이 외래어로 판단한 총 외래어 수}} \times 100$$

평가에 사용된 총 어절 수 : 19,000 어절
평가에 사용된 총 외래어 어절 수 : 4,709 어절
조건 : 최저음절 빈도수 - 2, 3, 5, 7, 10
평균음절 빈도수 - 2, 5, 6, 10, 13

4. 실험 결과 및 분석

총 25가지의 조건을 주고 성능을 평가하였는데 표에서 표현의 한계상 성능의 변화가 없는 부분은 제거하고 변화가 있는 부분만 표시를 하였다.

첫 번째 저빈도(외래어 어절 수 200미만) 일때의 경우를 보면 처음에는 인식율이 10~50%로 낮았지만 최저음절 빈도수와 평균음절 빈도수를 높일수록 35~85%까지 크게 오르는 것을 볼 수 있다. 정확률은 3가지 실험중에서 가장 큰 성능폭이 나타났다. 반면에 재현율은 처음 70%대에서 마지막에 30~40%대로 떨어지는 것으로 나타났는데 이는 정확률을 높이면 재현율이 떨어지는 trade-off 관계에 있기 때문이다. 두 번째 실험군인 평균적인 외래어 빈도수를 가지고 있는 문서(200~400미만)들의 정확률을 성능 평가를 보면 저빈도일 때보다 성능향상 폭이 그리 크지 않은 것을 볼 수 있고 역시 재현율은 큰 폭으로 감소하고 있는 것을 알 수 있다. 마지막 세 번째 실험군인 고빈도 외래어의 정확률의 경우는 조건을 변화시켜도 그리 변화가 없었다. 반면에 재현율의 경우는 20~30%정도의 감소가 있었다.

성능이 대체적으로 좋지 않은 경우는 중국 무협 소설이나 일본 게임에 대한 분야에서는 영어 외래어 자체가 별로 쓰이지 않아서 제대로 추출할 수가 없었다. 이는 중국 무협 소설의 경우 영어보다는 중국어가 많이 쓰이고 일본 게임 같은 경우는 일본어가 많이 쓰여서 아직 영어 이외의 외래어에 대해서는 적용하지 않고 있기 때문이다.

인식율이 좋지 못했던 경우는 한국의 판광지를 소개한 웹문서의 경우도 그리 좋지 못한 성능을 보였는데 ‘석가탑’, ‘청량리’ 등과 같이 외래어에서 많이 출현하는 음절들이 한국의 지명이름이나 마을이름, 산 이름 등에 많이 쓰이기 때문으로 생각된다. 또 성능이 좋지 않은 경우는 과학분야에서 특히 지질학이나 물리학 분야에서 많이 좋은 않은 성능을 보였다. 그 이유는 그 외국 이름을 그대로 쓰기보다는 우리말로 해석해서 많이 쓰이기 때문으로 풀이된다.

좋은 인식율을 보이는 문서는 컴퓨터 관련분야의 신문 기사나 뉴스 같은 문서, 게임사이트에서 추출한 문서 등으로 우리나라 게임이던 외국 게임이던지 상관없이 성능이 잘 나왔다. 이는 게임 상에 우리나라 말보다는 외국말들을 많이 채용해서 그렇기 때문으로 풀이된다. 또, 한국 사람 이름이 많이 나오는 경우의 성능도 좋았다. 그 이유는 한국 사람 이름은 한자나 한글 이름이 많기 때문에 널리 쓰이는 글자들이 많이 쓰이기 때문으로 분석된다.

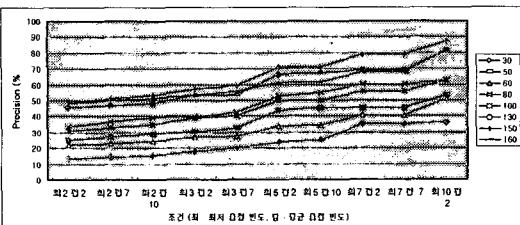


그림 2. 200개 미만의 외래어를 가지고 있는 문서들에 대한 정확률

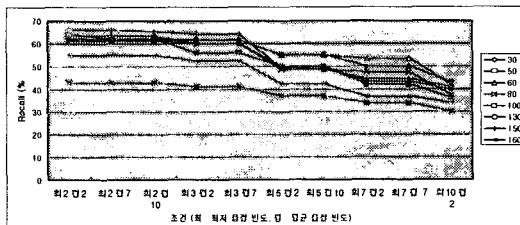


그림 3. 200개 미만의 외래어를 가지고 있는 문서들에 대한 재현율

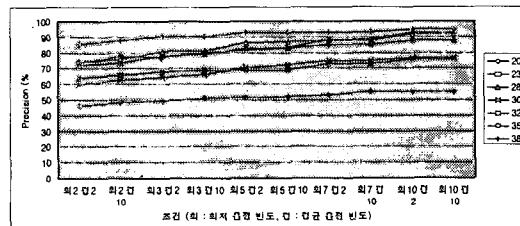


그림 4. 200~400개의 외래어를 가지고 있는 문서들에 대한 정확률

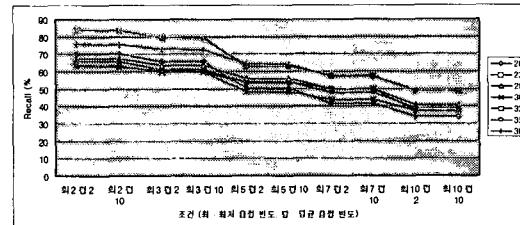


그림 5. 200~400개의 외래어를 가지고 있는 문서들에 대한 재현율

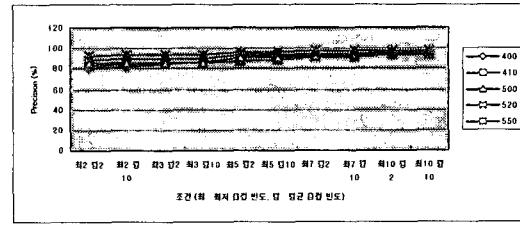


그림 6. 400개 이상의 외래어를 가지고 있는 문서들에 대한 정확률

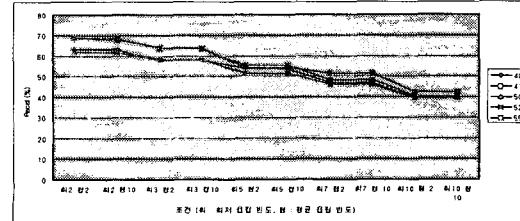


그림 7. 400개 이상의 외래어를 가지고 있는 문서들에 대한 재현율

5. 결론 및 향후연구

정보 검색에서 검색하는 단어의 사용에 있어서 올바른 단어들을 사용하면 사용자가 원하는 정확한 결과를 얻을 수 있지만 적절하지 못하거나 올바르지 못한 단어들을 쓰는 경우에는 사용자가 원하는 정보를 정확하게 얻기가 힘들다. 검색엔진에서 사용자가 입력한 검색어를 외래어인지 아닌지 판단해서 정확한 검색 결과를 보여준다면 사용자에게 좀더 나은 신뢰성과 검색 시간의 절약으로 더욱 빠른 검색 결과를 보여줄 수 있을 것이다.

본 논문에서는 외래어를 판단하는데 있어서 외래어 사전의 등록 없이 외래어의 음절 빈도수만을 이용한 시스템을 구현하였다. 웹에서 추출한 외래어로 외래어 표를 구성하고 최저 음절 빈도수와 평균 음절 빈도수를 조건으로 주었다. 일단 좋은 성능을 나타내는 IT분야나 과학분야에서 활용할 수 있을 것으로 본다. 여러 분야의 문서에서는 좋지 않은 성능을 보였는데 이는 단순히 외래어에 쓰인 글자들로만 판단하는 것이 아직은 미흡하다는 것을 보이는 것 같다. 향후 계획으로 아직 외래어 표에 자모빈도 규칙에 대한 것을 적용하지 못하는데 여기서 조사한 자모빈도를 바탕으로 조성, 증성, 종성에서의 가중치를 부여하여 서로 비교해볼 예정이다. 또, 현재는 띄어쓰기(복합명사분해)를 전혀 고려하지 않아서 '영동메론', '포도쥬스'와 같은 한글과 외래어가 같이 들어있는 단어들이 외래어로 분류가 되질 않는 문제가 발생하였고 단어의 길이 또한 고려하지 않아서 3음절 이하에서 많은 오류를 보이고 있으므로 띄어쓰기와 단어의 길이를 고려한다면 좀더 좋은 성능을 보일 수 있을 것이다. 마지막으로 현재는 영어에 관련된 외래어만 적용하고 있는데 다른 언어에 관련된 외래어도 추가한다면 성능이 더욱 좋아질 것으로 본다.

[참고문헌]

- [1] 문화관광부, 통신 언어 어휘집, 연구보고서, 2001.
- [2] 강승식, 한국어 형태소 분석과 정보 검색, 홍릉출판사, 2002.
- [3] 김영택 외 공저, 자연언어처리, 2001.
- [4] 김명철, 김덕봉, 김유성, 김재훈, 박혁로, 이하규, 최신 정보검색론, 2001.
- [5] 김진호, 류근호 공역, 정보 검색, 1995.
- [6] 박숙희, 뜻도 모르고 자주 쓰는 우리말 500가지, http://my.dreamwiz.com/yimdream/urimal/book1_1a.htm
- [7] 박성배, 장병탁, "음절 정보만 이용한 한국어 복합명사 분해", 제 15회 한글 및 한국어 정보처리 학회, pp.33-39, 2003.
- [8] 김용균, 서영훈, "미등록어 처리가 강화된 복합명사 분해", 제 15회 한글 및 한국어 정보처리 학술대회, pp.41-44, 2003.
- [9] 이재성, "이중언어 코퍼스로부터 외래어표기 사전의 자동구축", 제 11회 한글 및 한국어 정보처리 학술대회, pp.142-149, 1999.
- [10] 오종훈, 최기선, "온너마르코프 모델(HMM)을 이용한 과학기술문서에의 외래어 추출 모델", 제 11회 한글 및 한국어 정보처리 학술대회, pp.137-141, 1999.
- [11] 김선희, 윤준태, 송만석, "한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석", 정보과학회논문지 : 소프트웨어 및 응용 제 29권 제 6 호 (2002.6), pp.407-415, 2002.
- [12] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", 정보과학회논문지 : 소프트웨어 및 응용 제 29권 제 4 호 (2002.4), pp.258-269, 2002.