심 광 섭 교수지도

석사학위청구논문

SVM을 이용한 외래어 인식

2006

성신여자대학교 대학원

전산학과

권 미 영

SVM을 이용한 외래어 인식

심 광 섭 교수지도

이 논문을 석사학위 논문으로 제출함.

2005년 11월

성신여자대학교 대학원

전산학과

권 미 영

인 준 서

권미영의 석사학위 논문을 인준함.

심사위원_____인

심사위원_____인

심사위원_____인

성신여자대학교 대학원

나와 함께 하시는 하나님께

이 논문을 바칩니다.

논문개요

한국어 텍스트에서 발견되는 외래어의 수는 점점 증가하는 추세에 있다. 외래어는 대체로 고유명사나 전문용어로, 생산적인 어휘 유형이어서 미등록어 문제를 일으키며, 음차 표기 또한 단일하지 않아서 정보검색에서 색인어 불일치 문제를 일으켜 재현율에 영향을 미치고 있다.

따라서, 본 논문에서는 SVM을 사용하여 외래어를 인식하는 방법을 제시한다. 외래어 인식 문제는 외래어와 순수 한국어 명사의 분류로 재정의하였다. 음절 정보와 음소 정보, 선별된 음소 정보와 선별된 음절 정보를 자질 벡터 생성에 사용하며, 학습 자질 벡터 9000개에 대해 SVM 학습을 수행하고, 테스트 자질 벡터 1000개에 대해 SVM 분류를 수행한다.

평가 결과, 벡터 생성에 반영되는 정보에 따라 정밀도 88.65%, 정확도 90.69%, 재현율 86.14%, F-measure(β=1) 88.35를 갖는 베이스라인에 비해 정밀도 약 2-5%, 정확도 약 3-6%, 재현율 약 0.5-3%, F-measure 약 1.5-4.5의 성능향상을 보여주었다. 가장 좋은 성능을 보여준 실험은 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 반영하여 자질 벡터를 생성한 실험으로 10-fold cross-validation 테스트에서 정밀도 93.06%, 정확도 96.55%, 재현율 89.30%, F-measure(β=1) 92.78을 나타냈다.

목 차

논문개요

Ⅰ. 서론 ····· 1

Ⅱ. 관련 연구 ····· 3
1. 한국어정보처리에서 외래어 관련 연구 ····· 3
2. SVM ····· 6

Ⅲ. SVM을 이용한 외래어 인식 ····· 13
1. 자질 선택과 표현 ····· 13
2. 실험 데이터 구성 ····· 16
3. 학습 및 분류 과정 ····· 17

Ⅳ. 실험 및 평가 ····· 20
1. 평가 척도 ····· 20
2. 평가 방법 ····· 21
3. 베이스라인 설정 ····· 22
4. 성능 평가 ····· 23
4.1 학습 데이터 크기에 따른 성능 비교 ····· 23
4.2 자질 선택에 따른 성능 비교 ····· 24
5. 결과 분석 ····· 28

Ⅴ. 결론 및 향후 과제 ····· 31

참고문헌

ABSTRACT

그림 목차

2.1 최대 margin을 갖는 경계면 ····· 7
2.2 margin과 경계면의 수학적 표현 ····· 9
2.3 SVM의 기본 원리 ····· 10
2.4 SVM의 구조 ····· 12
3.1 SVM 입력 파일 포맷의 예 ····· 16
3.2 SVM을 이용한 외래어 인식에서 학습 과정 ····· 17
3.3 SVM을 이용한 외래어 인식에서 분류 과정 ····· 18

표 목차

3.1 외래어와 한국어에서의 각 음소 분포 현황 ····· 14
4.1 이진 결정에 대한 분할표 ····· 20
4.2 한자로 변환가능한 음절 개수를 이용한 분류의 성능 ····· 22
4.3 학습 데이터 크기에 따른 성능 비교 실험 결과 ····· 23
4.4 자질 벡터 크기에 따른 성능 비교 실험 결과 ····· 24
4.5 음절 정보 자질 선택과 자질 표현 방법에 따른 성능 비교 실험 결과25
4.6 자질 값에 반영되는 정보에 따른 성능 비교 실험 결과 ····· 26
4.7 자질 생성에 반영되는 정보에 따른 성능 비교 실험 결과 ····· 27
4.8 음절 정보와 선별된 음소·음절 정보를 사용한 실험의 집합별 결과 · 28
4.9 9번과 10번 집합에서 종류별로 오류 예제가 차지하는 비율 ····· 29
4.10 9번과 10번 집합의 오류 예제 ····· 29

Ⅰ. 서론

현대 한국어 텍스트에서 발견되는 외래어의 수는 점점 증가하는 추세에 있으며, 과학·공학 분야와 같은 전문 영역에서는 한글 대비 외래어의 비중이 압도적으로 더 높은 경우들도 발견된다. 외래어는 한국어 문장에서 명사 성분으로 사용되고, 대체로 고유명사나 전문용어가 많다. 음차표기 또한 단일하지 않아서 정보처리에서 다음과 같은 문제를 야기한다[3].

첫째, 외래어는 사전에 등재되지 않는 경우가 많아 한국어 형태소 분석에서 미등록어 문제의 원인이 된다. '오페라는'과 같은 어절에서 '오페라'가 미등록어일 경우에 두 가지 가능한 기능어는 '는'과 '라는'인데, 기능어의 최장일치에 의해 '라는'이 기능어로 처리되어 '오페라+는'으로 분석되어야 할 어절이 '오페+라는'으로 잘못 분석된다[5]. 둘째, 확률적 품사 태깅에서도 미등록어인 외래어가 문장에 포함되어 있으면 품사 조합 수가 늘어나므로 가장 좋은 품사열을 선택하는 데에 어려움이 생긴다. 셋째, 외래어는 정보 검색의 색인어로 자주 사용되는데, '디지탈', '디지틀', '디지털'처럼 음차표기가 다양한 외래어에 대해 색인어 불일치 문제를 일으킨다. 검색엔진에서 외래어 색인어 처리 문제는 정보 검색의 재현율을 높이는데 매우 중요한 요인이 된다[16].

이러한 문제를 해결하기 위해서 외래어 인식 문제를 다루는 연구가 필요하다. 본 논문은 외래어 인식 문제를 외래어와 순수 한국어 명사의 분류 문제로 정의하고, 이진 분류에 좋은 성능을 보이는 SVM(Support Vector Machines) 학습 방법을 이 문제에 적용하는 방안을 제시한다.

본 논문은 총 5장으로 구성되어있다. 2장에서는 한국어정보처리에서 외래어

문제를 다룬 기존의 연구에 대해 고찰하고, 이 실험에 사용되는 SVM에 대해서 설명한다. 3장에서는 외래어 인식에 중요한 영향을 미치는 자질 선택 및 표현에 중점을 두어 살펴보고, 실험 데이터 구성과 SVM 학습 및 분류 과정을 소개한다. 본 논문에서 제안한 SVM 학습 및 분류 과정을 가지고 여러 가지 주제로 성능 비교 실험을 수행했는데 4장에서는 그 내용을 설명하고, 그에 따른 결과를 분석한다. 마지막으로 5장에서는 결론 및 향후 과제를 제시한다.

Ⅱ. 관련 연구

1. 한국어정보처리에서 외래어 관련 연구

한국어정보처리에서 외래어를 다루는 연구 주제로는 외래어 인식 및 추출, 추출된 외래어에 대한 동등 클래스(equivalence class) 구성, 외래어와 원어 비교, 영-한 자동 음차표기 및 원어 복원 등이 있다.

Kwon[16]은 외래어 인식 문제를 주어진 어절에 외래어가 포함되어 있는지 없는지를 판단하는 문제로 정의하였고, 외래어 추출은 외래어 인식을 통해 외래어가 존재하는 것으로 판단된 어절에 포함된 외래어를 추출하는 문제로 정의하였다. 이 연구에서는 학습 코퍼스로부터 유니그램(unigram) 및 바이그램(bigram) 통계 조합을 구하여 외래어와 순수 한국어에 나타나는 음절 분포 차이에 대한 통계적 정보를 얻는 알고리즘을 제안하고, 이 알고리즘으로 주어진 어절에 외래어를 포함하는지 여부를 결정한다. 어절 W에 대해서 식(1)의 값이 1보다 크면 W가 외래어를 포함하는 것으로 판단한다. 이 식에서 $P(Foreign)$과 $P(Korean)$은 학습 코퍼스에서 외래어의 비율과 순수 한국어의 비율을 계산하여 추정하고, $P(W|Foreign)$과 $P(W|Korean)$은 단어에 나타난 음절의 유니그램과 바이그램을 이용하여 추정한다.

$$D(W) = \frac{P(Foreign|W)}{P(Korean|W)} = \frac{P(W|Foreign) \times P(Foreign)}{P(W|Korean) \times P(Korean)} \quad (1)$$

이 연구에서 제안한 알고리즘의 성능을 평가하기 위하여 컴퓨터·정보과학 분야의 문서로 구성된 KT 실험집합과 순수과학 분야의 문서로 구성된 KRIST 실험집합이 사용되었다. 실험 결과 KT 실험집합에서 정확률 97.1%, 재현율 66.78%를 기록하였고, KRIST 실험집합에서는 정확률 92.6%, 재현율 64.1%를 기록하였다. 또한, KT와 KRIST 실험집합을 통합한 실험에서는 정확률 96.6%, 재현율 64.6%의 결과를 얻을 수 있었다.

오종훈[5]의 경우, 외래어 인식 및 추출 문제를 음절태깅이라는 문제로 변환하여 해결하였다. 이 연구에서는 순수 한국어 음절 표현 상태와 외래어 음절 표현 상태의 이진 상태로 모델링한 은닉 마르코프 모델(Hidden Markov Model)을 이용하였다. 은닉 마르코프 모델에서는 학습 코퍼스로부터 추정된 전이확률, 음절확률, 자음확률을 사용하였다. 이 연구에서 제안된 방법을 KT 실험집합과 KRIST 실험집합을 대상으로 여러 가지 비교실험을 한 결과 기존 연구[16]에 비하여 높은 재현율과 정확률을 보였고, 자음정보의 유용성과 적은 양의 학습 코퍼스로도 좋은 성능을 나타냄을 보였다. 또한 이 기법은 실험집합이 같은 경우뿐만 아니라 실험집합이 다른 경우에도 좋은 성능을 나타냈다. 유형별 외래어 추출 실험 결과를 소개하면, 순수 한국어의 경우 재현율과 정확률 모두 약 99%를 보였고, 순수 외래어의 경우 실험 집합에 따라 재현율은 약 92-97%, 정확률은 약 98%를 보였다. 외래어와 한국어의 조합 유형은 실험 집합에 따라 재현율이 약 87-89%, 정확률이 약 84-86%를 보였다.

강승식[1]의 연구에서는 별도의 외래어 사전 없이 외래어의 음절 빈도수만을 이용하여 외래어를 판단하는 시스템을 구현하였다. 이 시스템에서는 웹에서 추출한 외래어로 외래어 표를 구성하고, 웹 문서에서 형태소 분석기를 통하여 추출한 단어를 외래어표와 한 음절씩 비교하여 모든 음절이 외래어 표에 있는 음절과 일치하면 그 단어는 외래어라고 1차 판정하였다. 외래어로 판

정된 단어의 각 음절 빈도수에 대하여 최저 음절 빈도수, 최대 음절 빈도수, 평균 음절 빈도수를 구한 후에 최저 음절 빈도수와 평균 음절 빈도수를 조건으로 주어서 최종적으로 외래어를 판단하는 실험을 실시하였다. 실험 결과 IT 분야나 과학 분야에서는 좋은 성능을 보였으나 다른 분야의 문서에서는 좋지 않은 성능을 보였다.

외래어의 동등 클래스를 구성하는 것은 원어에서 같은 단어인데 표기상 다른 외래어들을 하나의 클래스로 묶는 것이다. 이 주제에 대해서 Jeong[9]이 제안한 방법은 바이그램에 기반한 유사성(similarity) 계산과 Damerau-Levenshtein Metric(DLM)을 사용하는 것이다. 음절수가 같은 두 외래어에 대해서는 알파벳 수준에서 바이그램 비교를 하고, 유사성을 계산하여 일정한 값을 넘으면 대응되는 동등 클래스에 할당된다. DLM은 자동 철자 오류 교정에 사용되는 방법 중에 하나로, 두 단어 사이에 삽입, 삭제, 치환, 전위가 발생하는 최소의 빈도수를 측정한다. 두 문자열의 차이를 나타내는 Minimum Cost Function을 통해 DL계수를 구하고, DL계수가 3을 넘으면 두 단어는 동등 클래스에 속하는 것으로 판단한다. KT 실험집합과 KRIST 실험집합에 대해서 실험한 결과, 사전적(lexical)으로 볼 때 KT 실험집합에서 정확률 46.3%, 재현율 73.8%를 기록했고, KRIST 실험집합에서는 정확률 54.2%, 재현율 86.7%를 기록했다. 의미적(semantic)인 측면에서는 KT 실험집합에서 정확률 78.9%, 재현율 74.8%, KRIST 실험집합에서 정확률 87.5%, 재현율 83.3%를 나타냈다.

외래어와 원어 비교 문제에서는 원어를 자동으로 음차 표기한 후 이를 외래어와 직접 비교하는 방법과 공통의 기호체계로 바꾼 후에 외래어와 원어를 비교하는 방법이 사용되었다. 한국어 문서에서는 외래어와 영어를 비교하는 연구가 이루어졌는데, 이는 정보검색에서 다양하게 사용되는 외래어 색인어를

찾기 위함이었다[7].

영-한 자동 음차표기에는 피봇방식과 직접방식이 있다. 피봇방식에서는 영어를 외래어로 표기하기 위하여 영어 단어를 발음기호로 표기한 후 이를 다시 규칙에 따라 한국어로 음차표기 한다. 즉, 피봇방식은 영어자소를 음소로, 음소를 한글자소로 표기하는 방식이다. 직접방식은 영어자소를 문맥에 따라서 한글자소로 변환하는 방식이다[6][7]. 자소 단위의 정확도가 직접방식에서는 71.7%, 피봇방식에서는 70.6%, 혼합방식에서는 75.8%로 혼합방식이 가장 우수한 결과를 보였다. 여기서 직접방식은 다른 방식에 비해 처리가 쉽고, 역으로 변환하면 외래어 복원에 쉽게 사용될 수 있으므로 많이 사용된다[7]. 영-한 자동 음차표기에 관련된 최근 연구로는 한국어 번역문과 영어 원문으로 구성된 병렬 코퍼스로부터 자동으로 외래어 표기 사전을 구축하는 시스템을 제안한 이재성[7]의 연구가 있다. 실험 결과 수작업으로 전처리를 한 모델 중 가장 성능이 높은 것은 정확률 91.0%, 재현율 85.4%를 보였고, 모든 과정을 자동으로 한 모델 중에서 가장 성능이 높은 것은 정확률 68.3%, 재현율 68.3%를 보였다. 또 오종훈[6]의 연구에서는 음차표기를 자소의 음성적 변환 작업으로 보고 자소정보뿐만 아니라 음소정보를 이용한 음차표기법을 제안하였다. 이 연구에서는 약 60%의 단어 정확도(word accuracy)를 나타내었다.

2. SVM

SVM(Support Vector Machines)은 데이터를 positive class와 negative class로 나누는 이진 분류기(binary classifer)로써 상대 클래스와 가까운 거리에 위치하면서 해당 클래스의 경계를 이루는 데이터 집합 사이의 거리를

최대화하는 경계면을 찾음으로써 분류의 성능을 높이는 방법이다. 그림 3.1에서 점선은 두 클래스의 경계선을 가리키며 실선은 두 클래스를 분류하는 경계면이다. 분포된 각 클래스의 데이터에 도달하기까지 평행 이동하는 두 경계선의 폭을 margin이라고 하며, 최대 margin을 갖는 두 점선 상의 데이터를 support vector라고 한다. SVM 학습은 최대 margin을 갖는 최적의 결정 함수를 찾는 것이다[8].

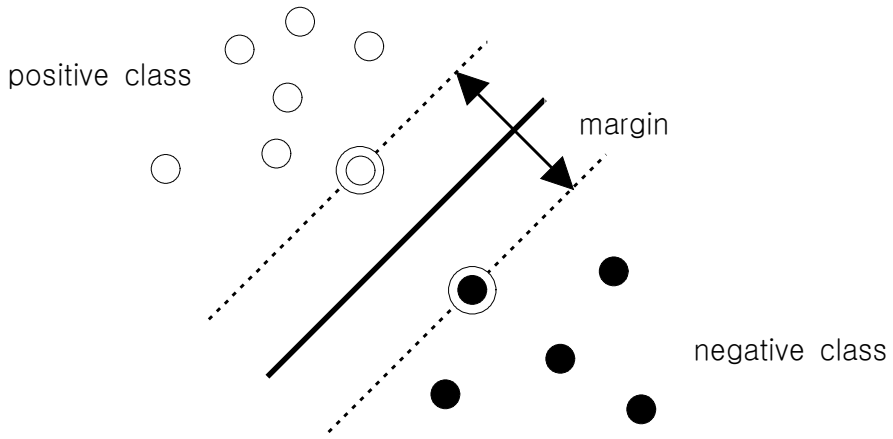

그림 2.1 최대 margin을 갖는 경계면

최적의 결정 함수를 찾기 위해 입력되는 데이터, 출력 결과, 클래스의 경계선과 margin을 수학적으로 표현하면 다음과 같다.

식(2)와 같이 입력 데이터 x_ℓ이 N차원의 벡터이고, 출력 결과 y_ℓ이 positive를 나타내는 +1 또는 negative를 나타내는 -1일 때,

$$(x_1, y_1), \ldots, (x_\ell, y_\ell) \in R^N \times \pm 1, \qquad (2)$$

두 클래스를 분류하는 경계면은 식(3)으로, 결정 함수는 식(4)로 표현할 수 있다.

$$(w \cdot x) + b = 0 \quad w \in R^N, b \in R, \qquad (3)$$

$$f(x) = sign((w \cdot x) + b) \qquad (4)$$

positive class의 경계선은 $(w \cdot x) + b = +1$, negative class의 경계선은 $(w \cdot x) + b = -1$이므로, positive class 영역은 $(w \cdot x) + b >= 1$, negative class 영역은 $(w \cdot x) + b <= -1$로 나타낼 수 있다.

margin을 구하기 위해서 negative class 경계선 상의 어떤 점 x^- 그리고, x^-와 가장 가까운 positive class 경계선 상의 점 x^+를 생각해보자. 어떤 값 λ에 대해서 $x^+ = x^- + \lambda w$ 이라고 할 수 있다. x^-에서 x^+까지의 직선은 경계선에 수직이고 w도 경계선에 수직이다. x^-에서 x^+까지의 직선을 얻으려면 w방향으로 얼마간 이동을 해야 한다. 지금까지 알고 있는 정보는 $(w \cdot x^+) + b = +1$, $(w \cdot x^-) + b = -1$, $x^+ = x^- + \lambda w$이다. $(w \cdot x^+) + b = +1$에 $x^+ = x^- + \lambda w$를 대입하여 풀면 $w \cdot x^- + b + \lambda w \cdot w = 1$이고, $(w \cdot x^-) + b = -1$이므로, $-1 + \lambda w \cdot w = 1$이다. 그러므로, $\lambda = \frac{2}{w \cdot w}$이다.

식(5)는 margin M을 구한 식이다.

$$M = |x^+ - x^-| = |\lambda w| = \lambda |w| = \lambda \sqrt{w \cdot w} = \frac{2\sqrt{w \cdot w}}{w \cdot w} = \frac{2}{\sqrt{w \cdot w}} \quad (5)$$

그림 3.2는 지금까지의 수학적 표현을 그림으로 정리한 것이다.

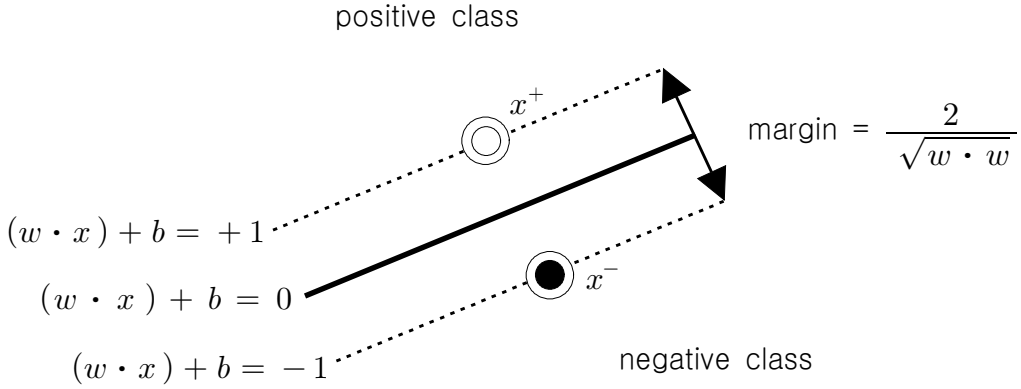

그림 2.2 margin과 경계면의 수학적 표현

이제 최대 margin을 구하는 문제는 $w \cdot w$를 최소화하는 이차 계획법(Quadratic Programming)[1] 문제로 풀 수 있다. 주어진 이차 계획법으로 구한 해 w는 margin 위에 있는 학습 데이터 집합을 사용해서 $w = \sum_i v_i x_i$로 나타

1) 이차 계획법(Quadratic Programming) : 엠파스 IT 용어사전에 따르면, 이차 계획법은 조건부 최대화 문제로서 목적 함수가 2차 함수, 제약 조건이 1차 부등식이나 등식으로 된 것. 비선형 계획법이라고도 한다. 선형 계획법(LP)과 같은 문제를 푸는 데 사용된다. LP는 1차 방정식으로 작성된 것을 사용하나 이것은 2차 방정식을 사용한다.

낼 수 있다. 여기서 v_i는 이차 계획법 문제를 해결함으로써 구할 수 있는 가중치이고, 학습 데이터 x_i는 support vector이다. support vector는 분류 문제를 푸는데 중요한 단서가 된다. 왜냐하면, 추출된 support vector를 제외한 모든 학습 데이터를 제거해도 동일한 결정 함수를 얻을 수 있기 때문이다.

w의 전개식을 이용해 최종 결정 함수를 식으로 나타내면 식(6)과 같다.

$$f(x) = sign(\sum_i v_i (x \cdot x_i) + b) \qquad (6)$$

SVM의 기본 원리는 그림 3.3과 같이 Φ를 통해 입력 공간의 학습 데이터를 dot product로 이루어진 고차원 자질 공간 F로 비선형적인 사상을 시키고,

$$\Phi : R^N \rightarrow F, \qquad (7)$$

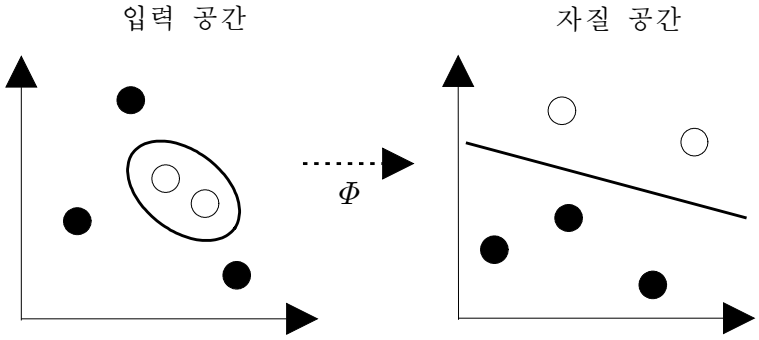

그림 2.3 SVM의 기본 원리

자질 공간 F에서 최대 margin을 가지는 경계면을 찾는 선형 알고리즘을 수행하는 것이다. 커널 함수를 사용하면 자질 공간으로의 사상을 명시적으로 수행하지 않고 최적의 경계면을 계산하는 것이 가능하다[10].

$$k(x, y) := (\Phi(x) \cdot \Phi(y)) \tag{8}$$

주로 많이 사용되는 커널 함수는 식(9)의 다항식 커널(polynomial kernel)과 식(10)의 RBF 커널(Radial Basis Function kernel), 식(11)의 시그모이드 커널(sigmoid kernel)이 있다.

$$k(x, y) = (x \cdot y)^d \tag{9}$$

$$k(x, y) = exp(- \mid x - y \mid^2 / (2\alpha^2)) \tag{10}$$

$$k(x, y) = tanh(\kappa(x \cdot y) + \Theta) \tag{11}$$

입력 공간에서 경계면은 비선형 결정 함수로 대응되는데 그 함수의 모양은 커널에 의해 결정된다. 커널 함수를 통한 최종 결정 함수는 식(12)와 같다.

$$f(x) = sign(\sum_{i=1}^{n} v_i \cdot k(x, x_i) + b) \tag{12}$$

그림 3.4는 SVM의 구조를 나타낸다. 테스트 벡터 x가 주어졌을 때 지금까지 설명한 계산 과정을 통해서 출력이 나오는 것을 보여주고 있다.

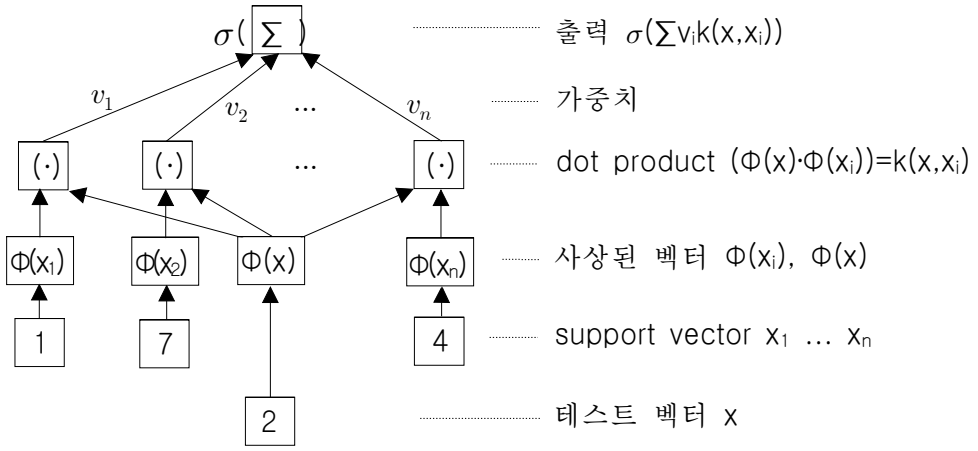

그림 2.4 SVM의 구조

이진 분류기인 SVM을 다중 분류기(multi-class classifier)로 확장하기 위한 방법에는 여러 가지가 있다. 그 중 가장 유명한 2가지 방법으로는 k개의 클래스가 있다고 가정할 때 k개의 분류기로 한 클래스와 그 외 다른 클래스로 분류하는 방법이 있고, 클래스의 모든 쌍을 고려한 k×(k-1)/2개의 분류기를 구축하여 가중치 투표로 최종 결정을 내리는 방법이 있다[13].

SVM은 초고차원의 학습 데이터를 가지고도 높은 일반화를 달성할 수 있고, 차원과 관계없이 작은 계산 오버헤드로 학습을 수행할 수 있다. 또한, 커널 함수의 도입으로 비선형 자질 공간을 다룰 수 있고 SVM을 도입한 실제 응용프로그램에서도 높은 성능을 보여주고 있다.

SVM이 활용되고 있는 대표적인 분야로는 문서 분류 및 범주화, 이미지 인식,손으로 쓴 문자 인식, 생물정보학에서 단백질 상동성 탐지와 유전자 표현 데이터의 자동 범주화 등이 있다[12].

본 연구에서는 순수 한국어 명사와 외래어를 분류하기 위해 Thorsten Joachims이 만든 SVMlight[14]을 사용한다.

Ⅲ. SVM을 이용한 외래어 인식

1. 자질 선택과 표현

외국어와 한국어는 음운학적인 차이가 있으므로 순수 한국어에서 자주 사용되는 음절과 외래어에서 자주 사용되는 음절은 서로 다를 것이다[5]. 또, 순수 한국어에서 자주 사용되지 않는 음소가 외래어에서는 자주 사용되는 것을 볼 수 있다. 'ㄹ, ㅋ, ㅌ, ㅍ, ㅐ, ㅡ'가 그 대표적인 예이다. 그래서, 순수 한국어 명사와 외래어를 분류하는데 중요한 단서를 제공하는 것을 음절과 음소로 간주하여 크게 네 가지 경우로 나누어 자질을 선택한다.

첫 번째로 음절 정보만을 자질로 선택하는 경우, 1부터 2350까지 KSC-5601 한글 영역 코드에 순서 번호를 준 값을 사용한다. 이 때, 자질 표현 방법에 따라 다시 세 가지로 나눌 수 있다. 단어에 나타난 음절을 순서대로 자질에 대응하여 자질을 표현하면 각 자질의 값은 대응된 음절의 KSC-5601 한글 영역 코드 순서 번호일 것이다. 이 표현 방법에는 음절 위치 정보와 음절 정보가 포함되며, 자질로 표현하고자 하는 음절수에 따라 자질의 수도 달라진다. 예를 들어, 단어의 최초 2음절을 자질로 선택한다면 자질의 수는 2개이며, 자질 값은 자질 위치에 대응하는 음절의 KSC-5601 한글 영역 코드 순서 번호가 될 것이다. 두 번째 표현 방법으로는 KSC-5601 한글 영역 코드를 자질로 하며 단어에 나타난 음절 빈도수를 자질 값으로 표현하는 것이다. 이 경우에는 음절의 위치 정보가 나타나지 않으며 자질의 수는 2350개다. 마지막으로, KSC-5601 한글 영역 코드를 자질로 하며 각 음절이 단어에 출현하는 여부를 자질 값으로 하는 경우가 있다.

두 번째로 음소 정보만을 자질로 선택하는 경우에는 KSSM 조합형 한글 코드에서 초성 30개, 중성 30개, 종성 30개, 총 90개를 자질로 하며, 해당되는 음소가 단어에 출현하는 빈도수를 자질 값으로 한다.

세 번째로 음절 정보와 선별된 음소 정보를 자질로 선택할 때는 KSC-5601 한글 영역 코드 2350개와 순수 한국어에는 자주 출현하지 않지만 외래어에는 자주 출현하는 음소 정보 6개, 총 2356개를 자질로 하며,

초성		중성		종성	
외래어	한국어	외래어	한국어	외래어	한국어
ㅇ : 3647	ㅇ : 1996	ㅣ : 4316	ㅏ : 2263	채움 : 15773	채움 : 4640
(0.088)	(0.080)	(0.104)	(0.090)	(0.381)	(0.185)
ㄹ : 2978	ㄱ : 1855	ㅏ : 3977	ㅗ : 2086	ㄴ : 1598	ㅇ : 2838
(0.072)	(0.074)	(0.096)	(0.083)	(0.039)	(0.113)
ㅅ : 2865	ㅅ : 1777	ㅡ : 3793	ㅣ : 2038	ㄹ : 1582	ㄴ : 2091
(0.069)	(0.071)	(0.092)	(0.081)	(0.038)	(0.083)
ㅌ : 1943	⋮	ㅗ : 2345	⋮	ㅇ : 661	ㄱ : 1120
(0.047)	ㄹ : 391	(0.057)	ㅡ : 410	(0.016)	(0.045
ㅍ : 1397	(0.016)	ㅐ : 2301	(0.016)	ㄱ : 402	ㅁ : 786
(0.034)	ㅍ : 215	(0.056)	⋮	(0.010)	(0.031)
ㅋ : 1394	(0.009)	⋮	ㅐ : 143	ㅁ : 395	ㄹ : 713
(0.034)	ㅌ : 149		(0.006)	(0.010)	(0.028)
ㅂ : 1230	(0.006)		⋮	ㅂ : 138	ㅂ : 201
(0.030)	⋮			(0.003)	(0.008)
⋮	ㅋ : 25			ㅅ : 73	ㅅ : 79
	(0.001)			(0.002)	(0.003)
	⋮			⋮	⋮

표 3.1 외래어와 한국어에서의 각 음소 분포 현황[2)]

2) 표 3.1에 각 음소별로 제시한 첫 번째 수는 해당 음소가 말뭉치에 출현한 빈도수를 나타내고, 괄호 안의 수는 해당 음소의 출현 빈도수를 전체 음절수로 나눈 값이다.

2350개의 자질에 대해서는 해당 음절이 단어에 출현하는 여부를, 6개 자질에 대해서는 해당 음소가 단어에 출현하는 빈도수를 자질 값으로 한다. 여기서 마지막 6개의 음소 정보는 순수 한국어 명사와 외래어 분류에 변별력 있는 음소를 통계적으로 선발한 것으로, 자음 정보 4개, 모음 정보 2개이다. 순수 한국어 명사, 외래어 각각 5000개를 대상으로 출현 빈도수와 $\frac{\text{각 음소의 출현 빈도수}}{\text{전체 음절 수}}$를 계산한 결과, 표 3.1과 같이 초성에서는 'ㄹ, ㅋ, ㅌ, ㅍ', 중성에서는 'ㅐ, ㅡ'가 순수 한국어와 외래어 간에 현격한 출현 차이를 보였으며, 종성에는 확연한 차이를 보여주는 음소가 없었다.

네 번째 경우는 세 번째 자질 집합에 선별된 403개의 변별력 있는 음절을 자질로 추가하여 총 2759개의 자질을 갖게 되며, 403개의 자질 값은 주어진 단어에서 해당 음절의 빈도수이다. 변별력 있는 음절을 선발하는 방법은 순수 한국어 명사 5000개와 외래어 5000개를 대상으로 KSC-5601 한글 영역 코드가 나타내는 각 음절의 출현 빈도수와 $\frac{ffreq}{ffreq+kfreq}\times 100$을 구해서 $\frac{ffreq}{ffreq+kfreq}\times 100$이 96.3 이상인 음절만 선택한다. 여기서 $ffreq$는 외래어 명사에서 해당 음절이 나타난 빈도수를, $kfreq$는 순수 한국어 명사에서 해당 음절이 나타난 빈도수를 의미한다.

위에서 제시한 자질 선택과 표현 방법을 바탕으로 학습 말뭉치로부터 입력 파일을 자동적으로 생성한다. 입력 파일은 '목표값 자질번호:자질값 자질번호:자질값...'의 포맷을 따르며 목표값은 외래어인 경우에는 1, 순수 한국어 명사인 경우에는 -1을 갖는다. 그림 3.5는 음절 정보 2350개와 음소 정보 6개를 자질로 하고, 음절 출현 여부와 음소 출현 빈도수를 각각의 자질 값으로 하여 학습 말뭉치로부터 입력 파일을 생성한 예이다.

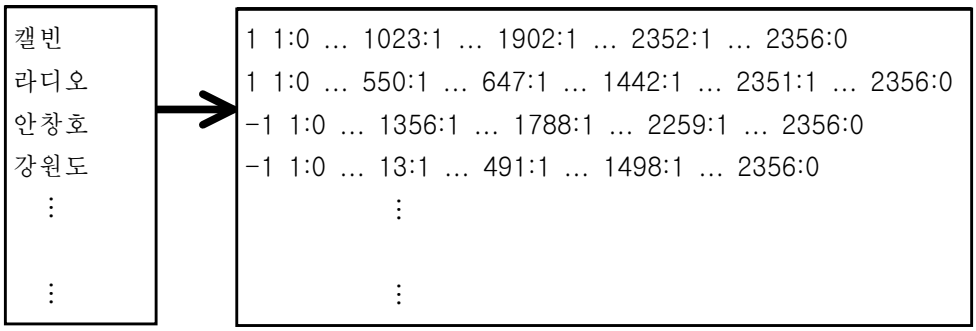

그림 3.1 SVM 입력 파일 포맷의 예

첫 번째 학습 데이터 '캘빈'은 외래어이므로 목표값 1을 출력하고, '캘'은 KSC-5601 한글 영역 코드 순서 번호가 1902이므로 1902번 자질에 1을 출력하며, '빈'은 KSC-5601 한글 영역 코드 순서 번호가 1023이므로 1023번 자질에 1을 출력한다. 이 단어에서 ㅋ의 출현 빈도수는 1이므로 'ㅋ'을 나타내는 2352번 자질에 1을 출력하고, 나머지 자질에는 모두 0을 출력한다.

2. 실험 데이터 구성

순수 한국어 명사와 외래어 실험 데이터는 각각 인명, 지명, 보통 명사로 분류하여 구성하였다. 외래어 실험 데이터는 국립국어연구원에서 2002년에 발간한 『외래어 표기 용례집』의 인명, 지명, 일반 용어부분에서 인명 1000개, 지명 1000개, 일반 용어 3000개를 뽑아 구성했다. 순수 한국어 인명 데이터는 KBS 인명사전에서 1000개를, 순수 한국어 지명 데이터는 국토지리정보원의 지형 지명 서비스에서 1000개를 뽑아 구성했다. 순수 한국어 일반 용어 데이터를 구축하기 위해서 국립국어연구원이 2003년 5월에 발표한 『한국

어 학습용 어휘 목록』에 포함된 보통 명사 중 2002년 국립국어연구원 보고서인 『현대 국어 사용 빈도 조사』의 빈도 순위를 기준으로 내림차순 정렬하여 상위 3000개를 선택했다. 이렇게 순수 한국어 명사 5000개, 외래어 5000개가 실험 데이터로 구성되었다.

3. 학습 및 분류 과정

SVM을 이용한 순수 한국어 명사와 외래어 분류에 대한 학습 과정은 그림 3.6과 같다. 우선 예제 구성 프로그램을 통해 순수 한국어 명사와 외래어 명사 말뭉치로부터 학습 예제 9000개와 테스트 예제 1000개를 구성한다. 이 말뭉치는 인명, 지명, 보통 명사로 구성된 순수 한국어 명사와 외래어 말뭉치이므로, 일정 비율로 각 카테고리의 예제를 뽑아 학습 예제와 테스트 예제를 구성하기 위해서 예제 구성 프로그램을 수행한다. 구성된 학습 예제는 말뭉치로

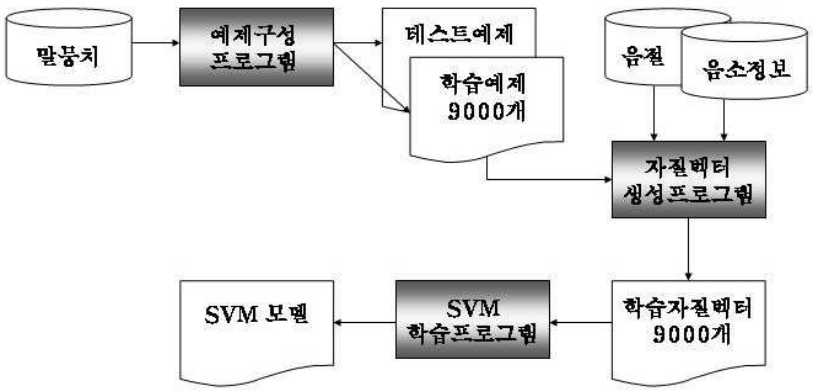

그림 3.2 SVM을 이용한 외래어 인식에서 학습 과정

부터 통계적으로 선발된 변별력 있는 음절 정보 및 음소 정보와 함께 자질 벡터 생성 프로그램에 입력으로 주어진다. 자질 벡터 생성 프로그램은 III장 1절에서 설명한 자질 선택과 표현 방법에 따라 학습 자질 벡터를 생성한다. SVM 학습 프로그램은 학습 자질 벡터 9000개를 입력으로 받아서 학습을 수행하고, 학습할 때 제시한 커널 파라미터 값과 support vectors, threshold b를 담은 SVM 모델을 출력한다.

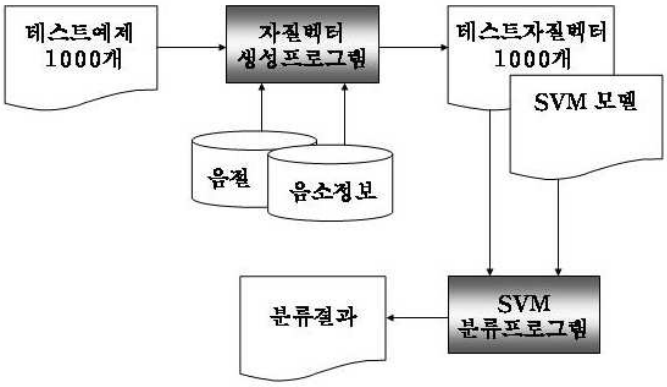

그림 3.3 SVM을 이용한 외래어 인식에서 분류 과정

분류 과정은 학습 과정에서 예제 구성 프로그램을 통해 구성된 테스트 예제 1000개와 SVM 학습 프로그램의 결과물인 SVM 모델을 사용한다. 학습 과정에서와 마찬가지로 테스트 예제 1000개에 대해 자질 벡터 생성 프로그램을 수행하여 테스트 자질 벡터 1000개를 얻는다. 테스트 자질 벡터는 학습 자질 벡터와 같은 포맷을 따르며, 분류 과정을 수행한 후에 얻게 되는 분류 결과와 비교하여 정확률(precision), 재현율(recall), 정밀도(accuracy)를 측정하기 위해 목표값도 함께 제시된다. 마지막 단계에서는 SVM 분류 프로그램이

테스트 자질 벡터 1000개와 SVM 모델을 입력으로 받아 분류를 수행하고 분류 결과를 출력한다. 외래어의 목표값은 1, 순수 한국어 명사의 목표값은 -1로 설정했기 때문에 분류 결과에 출력된 값이 양수이면 외래어, 음수이면 순수 한국어 명사로 분류된 것이다.

Ⅳ. 실험 및 평가

1. 평가 척도

이 실험에서는 평가 척도로 재현율(recall), 정확률(precision), 정밀도(accuracy), F-measure를 사용한다. 분류 결과가 표 4.1과 같을 때, 재현율, 정확률, 정밀도, F-measure를 구하는 식은 다음과 같다.

분 류 \ 정 답	외래어	순수 한국어	
외래어	a	b	a + b
순수 한국어	c	d	c + d
	a + c	b + d	a+b+c+d = n

표 4.1 이진 결정에 대한 분할표

■ 재현율 : $R = \frac{a}{a+c} = \frac{\text{외래어로 분류된 외래어의 개수}}{\text{외래어의 개수}}$

■ 정확률 : $P = \frac{a}{a+b} = \frac{\text{외래어로 분류된 외래어의 개수}}{\text{외래어로 분류된 단어의 개수}}$ (13)

■ 정밀도 : $Acc = \frac{a+d}{n} = \frac{\text{각각의 클래스로 올바르게 분류된 단어의 개수}}{\text{단어의 총 개수}}$

■ F-measure : $F_\beta(r,p) = \frac{(\beta^2+1)pr}{\beta^2 p + r}$ (r = 재현율, p = 정확률)

2. 평가 방법

이 실험에서는 10-fold cross-validation을 평가 방법으로 도입했다. k-fold cross-validation은 사용 가능한 예제의 수가 적을 경우 사용하는 방법 중에 하나이며 아래에 소개된 단계를 따른다.

- 전체 실험 데이터를 k개의 공통 원소를 가지지 않는 같은 크기의 집합으로 분할한다.

  $D_1, D_2, \cdots, D_k$

  $|D_i| = \frac{N}{k}$ (N : 전체 실험 데이터의 개수)

- 각각의 D_i에 대하여 다음 작업을 수행한다. $(i = 1, 2, \cdots, k)$
  - D_i를 제외한 나머지 $(k-1)$개의 집합으로 학습 집합을 구성
  - 학습 집합으로 학습 수행
  - D_i를 테스트 집합으로 하여 결과 측정
- 위에서 구한 모든 집합의 결과에 대해 평균을 구하여 이 실험의 최종 결과를 산출한다.

이 연구에서는 10000개의 예제에 대해서 10-fold cross-validation을 사용하므로, 10가지의 학습-테스트 집합이 구성되며, 각각의 학습 집합은 9000개의 예제로, 테스트 집합은 1000개의 예제로 구성된다.

3. 베이스라인 설정

실험을 통해 얻은 평가 척도의 수치만을 보고 실험 결과의 유의미성을 판단할 수는 없다. 주어진 문제의 난이도가 낮아서 평가 척도의 값이 대부분 높이 나오는 경우가 있는가 하면, 평가 척도의 값이 수치상으로는 낮지만 문제의 난이도에 비해서 상대적으로 높이 나오는 경우도 있다. 따라서, 이 실험에서는 베이스라인(baseline)을 구하고 그것을 기준으로 각 실험의 유의미성을 판단하기로 한다.

베이스라인은 주어진 문제를 풀 수 있는 단순한 알고리즘의 성능을 말한다. 베이스라인은 lower bound라고도 하며, 분류 문제를 다룰 때 lower bound와 upper bound는 실제로 중요한 기준이 된다.

순수 한국어 명사와 외래어 분류 문제에서 '순수 한국어 명사는 대부분 한자로 변환가능하다'는 것을 전제로 하고, 10000개의 실험 데이터에 대해서 두 가지 베이스라인을 구했다. 한자로 변환가능한 음절의 개수가 전체 음절의 개수와 같은 경우와 전체 음절의 개수-1과 같은 경우로 나누어 그 조건에 만족하면 순수 한국어 명사로, 그 외에는 외래어로 분류한다. 표 4.2는 두 가지 실험의 결과이다. 두 실험 중, 한자로 변환가능한 음절의 개수가 전체 음절의 개수와 같은 경우의 F-measure가 더 높게 나왔으므로 이 실험의 결과를 베이스라인으로 정했다.

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
단어 전체 음절	88.65	90.69	86.14	88.35
단어 전체 음절-1	77.63	97.16	56.92	71.78

표 4.2 한자로 변환가능한 음절 개수를 이용한 분류의 성능

4. 성능 평가

4.1 학습 데이터 크기에 따른 성능 비교

학습 데이터 크기에 따른 성능 비교 실험을 위해서 1000개의 데이터가 있는 말뭉치와 10000개의 데이터가 있는 말뭉치를 사용했다. 주어진 말뭉치에서 10%는 테스트 데이터로 하고 나머지는 학습 데이터로 사용하여 실험을 진행했다. 표 4.3의 비교 실험 결과를 살펴보면 900개의 학습 예제를 대상으로 실험한 것보다 9000개의 학습 예제를 대상으로 실험한 결과가 정밀도, 정확률, F-measure 모두 약 2-4% 높게 나왔다. 그런데 학습 데이터의 크기 증가에 따른 재현율은 자질의 개수가 2개인 첫 번째 실험에서 약 7%의 상승을 보였고, 자질의 개수가 2356개인 두 번째 실험에서 약 3%의 상승을 보였다.

▪ 자　질 : 단어의 최초 2음절
▪ 자질 값 : KSC-5601 한글 영역 코드 순서 번호

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
900개 학습예제	52.50	52.76	62.40	57.17
9000개 학습예제	56.20	54.62	69.06	61.00

▪ 자　질 : KSC-5601 한글 영역 음절 2350개, 음소 정보 6개
▪ 자질 값 : 2350개 음절의 단어 출현 여부를 1과 0으로 표시
6개 음소가 단어에 출현한 빈도수

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
900개 학습예제	90.30	94.34	86.20	90.09
9000개 학습예제	92.86	96.27	89.16	92.58

표 4.3 학습 데이터 크기에 따른 성능 비교 실험 결과

또한 표 4.3의 두 번째 실험 결과를 보면 900개의 학습 예제를 대상으로 한 것과 9000개의 학습 예제를 대상으로 한 실험 모두 F-measure가 90을 넘으므로, 학습 예제의 크기가 매우 작은 경우에도 상대적으로 좋은 성능을 나타내는 것을 알 수 있다.

4.2 자질 선택에 따른 성능 비교

첫 번째 실험은 음절 정보와 음절 위치 정보를 담고 있는 자질 벡터를 생성하되, 자질 벡터를 생성하는데 반영하는 음절의 길이를 달리한 경우에 대한 비교 실험이다. 단어의 최초 2음절만 자질 벡터 생성에 반영한 경우 벡터의 크기가 2이다. 단어의 전체 음절을 자질 벡터 생성에 반영한 경우에 SVM은 가장 긴 벡터의 크기를 기준으로 하여 전체 자질 벡터의 크기를 맞추므로, 학습 데이터에서 최장 단어 길이인 15가 벡터의 크기이다.

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
최초 2음절	56.20	54.62	69.06	61.00
전체 음절	77.15	95.91	56.72	71.28

표 4.4 자질 벡터 크기에 따른 성능 비교 실험 결과

표 4.4를 살펴보면 전체 음절을 자질 벡터 생성에 반영한 실험이 재현율은 떨어지지만 정밀도, 정확률, F-measure에서 큰 폭의 상승을 보였다. 이는 전체 음절을 사용한 경우가 단어 최초 2음절을 사용한 경우보다 더 많은 정보를 제공하기 때문인 것으로 볼 수 있다. 또 이 실험 결과는 순수 한국어 명사

와 외래어 분류 문제에 대한 베이스라인에 비해 매우 낮은 성능이지만 벡터의 크기가 성능에 큰 영향을 미침을 보여주고 있다.

두 번째 실험은 단어 전체에 대한 음절 정보를 사용하여 자질 벡터를 생성하되 자질 표현 방법을 달리한 경우에 대한 비교 실험이다. 첫 번째 경우, 자질 벡터의 원소는 주어진 단어의 각 음절에 대한 KSC-5601 한글 영역 코드의 순서 번호이다. 이 자질 벡터의 크기는 15이다. 다른 하나는 자질 벡터의 각 원소가 KSC-5601 한글 영역 코드의 각 글자를 가리키고 각 원소의 값은 해당 글자가 주어진 단어에 출현한 빈도수이며 벡터의 크기는 2350이다.

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
주어진 단어에 나타나는 모든 글자	77.15	95.91	56.72	71.28
KSC-5601 한글 영역 코드의 모든 글자	92.29	96.04	88.20	91.95

표 4.5 음절 정보 자질 선택과 자질 표현 방법에 따른 성능 비교 실험 결과

표 4.5는 실험 결과를 나타낸다. KSC-5601 한글 영역 코드의 각 글자를 자질 벡터의 원소로 하고 해당 글자가 주어진 단어에 출현한 빈도수를 자질 값으로 표현한 경우에 정밀도, 정확률, 재현율, F-measure 모두 월등히 높았으며, 베이스라인과 비교해 볼 때도 모든 부문에서 성능이 향상되었다. 이것은 자질 벡터 표현 방법이 성능에 많은 영향을 미침을 보여준다. 표 4.5로부터 같은 정보를 표현하더라도 자질의 수가 많은 방향으로 자질 벡터를 나타내는 것이 더 좋은 성능을 얻는데 도움이 되는 것으로 판단할 수 있다.

세 번째로 같은 자질을 선택하였으나 자질 값이 다른 정보를 가지는 경우를 비교해 보았다. 표 4.6은 KSC-5601 한글 영역 코드의 글자를 자질 벡터의

원소로 하고, 해당 글자가 단어에 출현한 빈도수를 자질 값으로 표현한 경우와 해당 글자가 단어에 출현한 여부를 자질 값으로 표현한 경우에 대해 실험한 결과이다. 실험 결과의 차이는 거의 없는데 이는 대부분의 음절이 한 단어에 한 번씩 나오기 때문인 것으로 보인다.

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
음절 출현 빈도수	92.29	96.04	88.20	91.95
음절 출현 여부	92.33	96.02	88.30	92.00

표 4.6 자질 값에 반영되는 정보에 따른 성능 비교 실험 결과

네 번째로 벡터 생성에 반영되는 정보에 따른 성능 비교 실험을 수행했다. 비교 대상은 벡터 생성에 음절 정보만 반영한 경우와 음소 정보만 반영한 경우, 음절 정보와 선별된 음소 정보를 반영한 경우, 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 반영한 경우이다. 음절 정보만 반영한 자질 벡터는 KSC-5601 한글 영역 코드의 각 글자를 원소로 하고, 해당 글자가 단어에 출현한 여부를 자질 값으로 표현했다. 음소 정보만 반영한 자질 벡터는 KSSM 한글 영역 코드에서 초성 30개, 중성 30개, 종성 30개를 원소로 하여 각 음소가 단어에 출현한 빈도수를 자질 값으로 표현한 벡터이다. 자질 값으로 출현 빈도수를 선택한 이유는 한 단어에 같은 음소가 여러번 나오기가 쉽기 때문이다. 음절 정보와 선별된 음소 정보를 반영한 자질 벡터는 음절 정보만 반영한 자질 벡터에 순수 한국어 명사와 외래어 분류에 변별력이 있는 자음 'ㄹ, ㅋ, ㅌ, ㅍ'와 모음 'ㅖ, ㅡ'를 추가한 벡터이다. 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 반영한 자질 벡터는 음절 정보와 선별된 음소 정보를 반영한 자질 벡터에 선별된 403개의 변별력 있는 음절을 추가한 벡터이다.

	정밀도(%)	정확률(%)	재현율(%)	F-measure(β=1)
음절 정보	92.33	96.02	88.30	92.00
음소 정보	90.34	93.63	86.50	89.92
음절 정보, 선별된 음소 정보	92.86	96.27	89.16	92.58
음절 정보, 선별된 음소·음절 정보	93.06	96.55	89.30	92.78

표 4.7 자질 생성에 반영되는 정보에 따른 성능 비교 실험 결과

실험 결과 표 4.7과 같이 음소 정보만을 반영한 것보다 음절 정보만 반영한 것이 더 좋은 성능을 보였고, 음절 정보와 선별된 음소 정보를 함께 반영한 것은 음절 정보만 반영한 것보다 모든 부문에서 소폭의 성능 향상을 보였다. 또한 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 반영한 벡터가 네 가지 경우의 벡터 중에서 가장 좋은 성능을 나타냈다. 따라서 음소 정보보다는 음절 정보가 외래어를 분류하는데 더 적합한 정보이며, 선별된 음소와 선별된 음절 정보가 성능 향상에 도움이 되는 정보라 할 수 있다. 그리고 선별된 음소 정보는 선별된 음절 정보보다 외래어 분류에 더 긍정적인 영향을 미친다고 할 수 있다.

5. 결과 분석

본 연구에서는 10-fold cross-validation을 도입하였으므로 10가지의 학습-테스트 집합에 대해서 실험 결과를 얻을 수 있었다. 대표적인 예로 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 사용한 실험의 집합별 결과를 표 4.8에 제시한다.

집합	SV의 수	정밀도(%)	정확률(%)	재현율(%)
1	1877	93.30	96.36	90.00
2	1947	96.00	97.13	94.80
3	1903	94.50	98.06	90.80
4	1859	92.50	96.30	88.40
5	1924	94.30	95.48	93.00
6	1904	94.90	97.66	92.00
7	1911	94.30	96.44	92.00
8	1874	93.40	97.17	89.40
9	1786	88.80	96.41	80.60
10	1784	88.60	94.47	82.00
평균	1876.9	93.06	96.55	89.3

표 4.8 음절 정보와 선별된 음소·음절 정보를 사용한 실험의 집합별 결과

앞에서 수행한 모든 실험 결과는 집합별로 표 4.8과 비슷한 양상을 보이고 있다. 모든 실험에서 9번과 10번 집합은 정밀도에서 90%를 넘지 못했으며, 재현율에서도 다른 집합과 현격한 차이를 나타냈다. 오류 분석을 위해 9번과 10번 집합의 결과를 분석한 결과가 표 4.9이다.

집합	분류	오류 예제가 차지하는 비율 (%)			
		인명	지명	보통명사	합계
9	외래어	3	59	11.6	19.4
	한국어	0	0	5	3
10	외래어	23	36	10.3	18
	한국어	2	0	7.3	4.8

표 4.9 9번과 10번 집합에서 종류별로 오류 예제가 차지하는 비율

표 4.9를 살펴보면 외래어인데 순수 한국어로 분류된 단어가 많았고 그 중에서 외래어 고유명사인 인명과 지명이 상당 부분 순수 한국어로 분류된 사실을 알 수 있었다. 반면, 순수 한국어를 외래어로 분류한 단어는 대부분 보통명사에 속했다.

외래어			순수 한국어	
인명	지명	보통명사	인명	보통명사
세모쥔	칭짱 공로	리포밍	홍난파	큰아들
쑤정핑	허난	우두	탄공	파란색
사오화쩌	허장 강	턱		모레
순친왕	비제	실		고춧가루
네얼	바안	쓰나미		코피
원이둬	하수	시밍		쓰레기통
쑨원	청청 산	수단		잔디
양중	추슝	위지위그		게시판

표 4.10 9번과 10번 집합의 오류 예제

외래어 인명과 지명 데이터에서 오류를 나타낸 단어는 중국어가 원어인 외래어였다. 이는 중국어의 음운체계가 서양어의 음운체계와 다르며, 학습한 데이터의 대부분이 서양어에서 비롯된 외래어이고, 실험에서 선택한 자질 중 선별된 음소 정보는 중국어보다는 서양어를 음차표기한 외래어와 관련된 특징을 반영하기 때문인 것으로 해석된다. 이 문제를 해결하기 위해서는 중국어를 음차표기한 외래어 고유의 특징을 반영하는 정보를 선택하여 자질 벡터를 생성하는 연구가 필요하다.

9번과 10번 테스트 집합의 외래어 지명 데이터와 보통명사 데이터에서 오류를 나타낸 단어 중에는 '실', '우한', '턱', '초', '수단', '둔화', '동경' 등과 같이 순수 한국어에서 쓰는 명사와 같은 음을 가진 외래어도 여러 개 있었다. 그 외에 오류를 나타낸 외래어와 순수 한국어 명사는 상대 클래스에서 자주 쓰이는 음절 또는 음소가 포함되어 있었다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 외래어 인식 문제를 외래어와 순수 한국어 명사의 분류 문제로 정의하고, SVM을 사용하여 이 문제를 해결하는 방법을 제시하였다. 음절 정보와 음소 정보, 선별된 음소 정보와 선별된 음절 정보를 추출하여 학습 자질 벡터 9000개와 테스트 자질 벡터 1000개를 생성하고, SVM 학습 프로그램을 통해 학습을 한 후, SVM 분류 프로그램으로 분류 결과를 얻었다. 평가 결과를 비교하기 위해 정밀도 88.65%, 정확도 90.69%, 재현율 86.14%, F-measure 88.35를 갖는 베이스라인을 구했다. 이 실험은 10-fold cross-validation 평가 방법을 사용하였다. 실험 결과, 벡터 생성에 반영되는 정보에 따라 베이스라인보다 정밀도 약 2-5%, 정확도 약 3-6%, 재현율 약 0.5-3%, F-measure 약 1.5-4.5%의 성능향상을 보여주었다. 가장 좋은 성능을 보여준 실험은 음절 정보와 선별된 음소 정보, 선별된 음절 정보를 반영하여 자질 벡터를 생성한 실험이었고, 정밀도 93.06%, 정확도 96.55%, 재현율 89.30%, F-measure(β=1) 92.78을 나타냈다.

학습 데이터 크기에 따른 성능 비교 실험에서는 크기가 매우 작은 학습 데이터를 사용해도 상대적으로 좋은 성능을 나타내는 것을 알 수 있었다. 또한 자질의 수가 많은 방향으로 자질 벡터를 표현하는 방법이 좋은 성능을 나타냄을 알 수 있었다. 음절 정보, 음소 정보, 선별된 음소 정보, 선별된 음절 정보를 사용한 비교 실험도 하였다. 4가지 정보를 모두 사용한 경우가 가장 좋은 성능을 보였고, 음소 정보만을 사용할 때 가장 낮은 성능을 보였다. 성능에 긍정적인 영향을 미친 정보를 순서대로 나열하면 2350개의 자질로 표현된 음절 정보, 6개의 선별된 음소 정보, 403개의 선별된 음절 정보이다.

오류를 분석하기 위해 가장 낮은 성능을 보인 9번과 10번 학습-테스트 집합을 선택하였다. 오류는 외래어에서 많이 나타났으며, 카테고리 별로 살펴볼 때, 외래어에서는 인명과 지명, 순수 한국어에서는 보통 명사에 오류가 집중되어 있음을 알 수 있었다. 9번과 10번 테스트 집합의 외래어 인명, 지명 부분은 중국어와 일본어를 음차표기한 외래어로 구성되었고, 오류는 중국어를 음차표기한 외래어에서 나타났다. 오류의 주된 원인은 학습 데이터의 대부분이 중국어와 음운체계가 다른 서양어에서 비롯된 외래어이며, 자질에 반영한 선별된 음소 정보는 서양어 관련 외래어의 특징을 나타내기 때문인 것으로 풀이된다. 또, 순수 한국어에서 쓰이는 명사와 같은 음을 가진 외래어와 상대 클래스에 자주 쓰이는 음절 또는 음소가 포함된 명사가 오류를 보였다.

따라서 향후 음운 체계가 비슷한 원어를 묶어서 외래어 분류에 변별력을 가지는 자질을 추출하는 연구가 진행되면 순수 한국어와 외래어 분류 성능이 향상될 것으로 보인다. 또한 원어에 따른 특징을 뽑아 자질 벡터를 생성한다면 원어별로 외래어를 분류할 수 있을 것이며, 그 결과를 가지고 외래어에 대응되는 원어를 추출하는 연구를 통해 표기가 통일되지 않은 외래어를 원어와 매치시킬 수 있다면 정보 검색 성능을 향상시키는 효과를 가져올 것이다.

현재 SVM으로 외래어를 인식하기 위해 음소 전이 정보를 자질에 반영하는 방법은 연구 중에 있다. ‘draft : 드래프트’의 경우, 밑줄 친 부분은 영어에서 ‘자음+자음’의 형태이다. ‘자음+자음’은 한국어에서 ‘종성+초성’일 때 나타나는 패턴이다. 그런데 이 부분을 음차표기한 글자는 ‘프트’로, 해당 알파벳이 일반적으로는 자음 ‘ㅍ, ㅌ’로 표기되는데 여기에 모음 ‘ㅡ’를 각각 결합한 것이다. ‘film : 필름’에서 알파벳 ‘l’은 한글로 ‘종성 ㄹ + 다음 음절 초성 ㄹ’로 표기된다. ‘idea : 아이디어’는 알파벳 모음 ‘i’가 한글로 2개의 모음 ‘ㅏ+ㅣ’으로 표기된다. 그 외에 외래어를 표기할 때 여러 가지 음소 전이 특징이 나타

난다. 따라서, 음소 전이 정보는 순수 한국어와 외래어를 분류하는데 긍정적인 영향을 미칠 것으로 기대된다. 단일 SVM을 사용할 때 음소 전이 정보를 자질에 반영하는 방법은 두 가지로 생각하고 있다. 하나는 단어에서 2음절씩 읽어서 2음소부터 6음소까지 구성 가능한 패턴을 생성하여 실험 데이터에서 통계적 방법을 이용해 유용한 음소 패턴 규칙을 뽑아서 자질로 표현하는 것으로, 현재 음소 패턴 규칙은 선별하였고 평가 단계만 남았다. 다른 한 방법은 2음소로 구성된 패턴이 실험 데이터에서 출현하는 확률을 계산하여 그 값을 저장하고, 단어에 대해서 확률을 적용하여 자질을 생성하는 방법으로 이 방법은 연구가 더 필요하다.

앞으로, 기존 연구에서 사용된 KT 실험집합과 KRIST 실험집합에 대해 이 논문에서 제시된 방법으로 분류 실험을 하여 기존의 연구 방법과 성능을 비교한 실험도 필요할 것이다.

참고문헌

[1] 강승식, 전영진, “음절 빈도를 이용한 외래어 명사의 인식”, 2003년도 한국멀티미디어학회 추계학술발표대회논문집, pp.408-411, 2003.

[2] 김영택 외, *자연언어처리*, 생능출판사, 2001.

[3] 남지순, “표기 중의성 외래어의 자동 처리를 위한 외래명사 전자사전의 구축”, 언어학, 제41권, pp.47-74, 2005.

[4] 문교부, “외래어 표기법”, 문교부, 1986.

[5] 오종훈, 최기선, “은닉 마르코프 모델을 이용한 음차표기된 외래어의 자동인식 및 추출 기법”, 인지과학, 제12권, 제3호, pp.19-28, 2001.

[6] 오종훈, 최기선, “자소 및 음소 정보를 이용한 영어-한국어 음차표기 모델”, 정보과학회논문지(B), 제32권, 제4호, 한국정보과학회, 2005.

[7] 이재성, “영-한 병렬 코퍼스로부터 외래어 표기 사전의 자동 구축”, 컴퓨터교육학회논문지, 제6권, 2호, pp.9-21, 한국컴퓨터교육학회, 2005.

[8] Andrew W. Moore, "Support Vector Machines", Tutorial Slides, School of Computer Science, Carnegie Mellon University, 2003.

[9] Kil-Soon Jeong, Yun-Hyung Kwon and Sung Hyun Myaeng, "Construction of Equivalence Classes of Foreign Words through Automatic Identification and Extraction", In *Proceedings of the NLPRS'97*, Tailand, 1997.

[10] Marti A. Hearst, "Trends and controversies : Support vector machines", *IEEE Intelligent Systems*, Vol.13, No.4, pp.18-28, 1998.

[11] Manning and Christopher D, *Foundations of statistical natural language processing*, MIT Press, 1999.

[12] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[13] T. Kudo and Y. Matsumoto, "Chunking with Support Vector Machines", In *Proceedings of the NAACL-2001*, 2001.

[14] Thorsten Joachims, "SVM-Light Support Vector Machine", available at http://svmlight.joachims.org/.

[15] Tom M. Mitchell, *MACHINE LEARNING*, McGraw-Hill INTERNATIONAL EDITIONS, 1997.

[16] Yun-Hyung Kwon, Kil-soon Jeong and Sung-Hyun Myaeng, "Foreign Word Identification Using a Statistical Method for Information Retrieval", In *Proceedings of the 17th International Conference on Computer Processing of Oriental Languages*, HongKong, 1997.

ABSTRACT

Foreign Words Identification Using Support Vector Machines

Kwon, Mi Young
Department of Computer Science
Graduate School of
Sungshin Women's University

Foreign words are often found in Korean texts. Most foreign words are proper nouns or technical terms, which are not in a dictionary. The variety of transliteration causes index term mismatch problem in Korean information retrieval, so that it influences recall of information retrieval.

This thesis proposes a SVM approach for foreign words identification in Korean texts. We consider the foreign words identification problem as a classification problem. Syllable information, phoneme information, selected phoneme information and selected syllable information are used in providing input vectors for SVM. 9000 training feature vectors are used for SVM learning and 1000 test feature vectors for classification by SVM.

Compared with the baseline, the proposed method improved the accuracy by 2-5%, the precision by 3-6%, the recall by 0.5-3%, and the F-measure by 1.5-4.5, depending on feature selection. The experiment with syllable information, phoneme information, selected phoneme information and selected syllable information showed the best performance. This experiment showed 93.06% accuracy, 96.55% precision, 89.30% recall and 92.78 F-measure(β=1) on 10-fold cross-validation tests.