Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 4, August 2015, pp. 386-391
http://dx.doi.org/10.5391/JKIIS.2015.25.4.386

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

한중 자동 문서분류를 위한 최적 자질어 비교

Comparison Between Optimal Features of Korean and Chinese for Text Classification

임미영* · 강신재**†
Mei-Ying Ren and Sinjae Kang†

*대구대학교 컴퓨터정보공학과, **대구대학교 컴퓨터·IT공학부
*Dept. of Computer & Information Engineering, Daegu University
**School of Computer & Information Technology, Daegu University

요 약

본 논문에서는 한국어와 중국어의 언어학적인 특징을 고려하여 문서 자동분류 시스템의 성능을 높일 수 있는 최적의 자질어 단위를 제안한다. 언어 종속적 단위인 형태소 자질어와 언어 독립적 단위인 n-gram 자질어 그리고 이들을 조합한 복합 자질어 집합을 대상으로 각 언어의 인터넷 신문기사를 SVM으로 분류하는 실험을 수행하였다. 실험 결과, 한국어 문서분류에서는 bi-gram이 F1-measure 87.07%로 가장 좋은 분류 성능을 보였고, 중국어 문서분류에서는 'uni-gram·명사·동사·형용사·사자성어'의 복합 자질어 집합이 F1-measure 82.79%로 가장 좋은 성능을 보였다.

키워드 : 중국어 문서분류, 한국어 문서분류, 정보획득량, SVM 분류기, 자질어 선택

Abstract

This paper proposed the optimal attributes for text classification based on Korean and Chinese linguistic features. The experiments committed to discover which is the best feature among n-grams which is known as language independent, morphemes that have language dependency and some other feature sets consisted with n-grams and morphemes showed best results. This paper used SVM classifier and Internet news for text classification. As a result, bi-gram was the best feature in Korean text categorization with the highest F1-Measure of 87.07%, and for Chinese document classification, 'uni-gram+noun+verb+adjective+idiom', which is the combined feature set, showed the best performance with the highest F1-Measure of 82.79%.

Key Words : Chinese Text Classification, Korean Text Classification, Information Gain, SVM Classifier, Feature Selection

Received: Mar. 3, 2015
Revised : Jun. 11, 2015
Accepted: Jun. 11, 2015
† Corresponding author
sjkang@daegu.ac.kr

1. 서 론

온라인 정보의 양이 날이 갈수록 늘어남에 따라 이러한 정보들을 자동적으로 처리하는 문서 자동 분류, 자동 요약, 질의응답시스템 등 지능 서비스의 필요성은 더욱더 커지고 있다.

하나의 문서를 전산처리하기 위해서는 프로그램 내부적으로 처리 가능한 형태로 변환하여 문서의 의미를 표현해야 하는데, 가장 대표적인 방법이 단어 목록(bag of words)이다. 여기서 단어(자질어)는 여러 형태를 가질 수 있으며 가장 대표적인 것이 언어 종속적인 자질인 형태소와 언어 독립적 자질인 n-gram이 있다.

한국어와 중국어 등 각 언어로 표현된 개별 문서뿐만 아니라 한 문서에 두 가지 언어가 동시에 출현하는 경우도 포함하여 처리할 수 있는 다국어 문서분류 시스템의 효과적인 구현을 위하여, 각 언어의 특성에 맞는 최적 자질어를 선정하고자 한다.

본 논문에서는 한국어 문서분류와 중국어 문서분류에서 어떤 자질어 집합이 가장 좋은 성능을 보이는지 실험을 진행하여 그 결과를 비교·분석하였다. 형태소 자질의 평가를 위해서 한국어 실험에서는 명사 집합, 명사·동사·형용사 집합을 선정하였고, 중국어 실험에서는 중국어가 가지고 있는 특수한 품사인 사자성어를 고려하여 명사 집합, 명사·동사·형용사·사자성어 집합 그리고 명사·사자성어 집합을 선정하였다. n-gram 자질의 평가를 위해서는 각

본 논문은 2013학년도 대구대학교 학술연구비에서 지원하여 연구하였음.

언어별 특징을 고려하여 한국어 실험에서는 bi-gram과 tri-gram을, 중국어 실험에서는 uni-gram과 띄어쓰기 없는 문장에서 음절 단위로 추출한 bi-gram, 그리고 형태소 분석을 통해 형태소 단위로 추출한 bi-gram 집합을 선정하였다.

그 다음으로 각 자질어 집합을 조합한 복합 자질어에 대한 실험도 수행하였는데, 기본적으로 uni-gram·bi-gram 조합을 선정하고, n-gram 자질어 실험과 형태소 자질어 실험에서 좋은 성능을 보인 것들을 대상으로 조합하여 복합 자질어 실험을 진행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 문서분류 관련 기존 연구들을 소개하고, 3장에서는 본 논문에서 제안하는 문서 분류 방법을 설명한다. 4장에서는 결과 분석을 기술하고, 5장에서 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

문서 분류 알고리즘에 대한 비교를 수행한 연구로는 [1,2,3,4] 등이 있는데, [1,2,3]에서는 SVM 분류기의 성능이 가장 좋은 것으로 나타났고, [4]에서는 비지도 학습과 SVM 분류기를 결합하여 자체 개발한 알고리즘이 가장 좋은 결과를 보였다. 또 [5]에서는 빈도수를 이용한 네트워크 분할 기법을 사용하여 마이크로블로그 글에 대하여 토픽 추출을 진행하였는데 효율적인 주제 발견 및 분할이 가능함을 확인하였다.

자질어 선정에 관한 연구 [6,7,8]에서는 정보획득량(information gain)과 카이제곱(chi-square)이 거의 비슷하게 좋은 것으로 나타났으며, [9]에서는 개선된 상호정보(mutual information) 방식이 가장 좋은 성능을 나타냈다. 그 외에 퍼지추론을 이용한 개선된 TF·IDF 기법을 자질 선택 방법으로 하여 소수 문서의 대표 키워드 추출을 진행한 연구[10]도 있었다.

영어권 자질어에 대한 관련 연구를 보면 [1,6]에서는 단어를 기본 자질어로 하였다. 이 중 [1]에서는 품사정보를 추가하여 성능향상을 가져왔고, 두 연구 모두 uni-gram과 bi-gram을 사용하였다. 그 외 [11]에서는 구문정보, 통사정보와 같은 언어학적 정보를 추가하였고, [12]는 워드 넷을 이용하였고, [13]에서는 어휘정보에 통사정보를 추가하여 성능을 제고하였다. 또 [14]에서는 감성 분석의 성능 제고를 위하여 SentiWordNet을 사용하였다. [15]는 자질어 중 n-gram에 대하여 통합적인 비교를 진행하였다. 각각 uni-gram, bi-gram과 Stanford parser를 이용한 triple과 AEGIR parser를 이용한 triple을 사용하였다. 그 결과는 네 가지 자질어를 모두 특징으로 사용하였을 때 성능이 가장 높게 나왔는데, 그 중에서도 bi-gram이 문서분류의 성능향상에 가장 많이 기여한 것으로 분석되었다.

한국어 문서분류 자질어에 관한 연구로는 [8,16]이 있었는데, 그 중 [8]은 특정 단어가 가지는 가중치는 시기에 따라 달라진다는 점을 이용하여 날짜 정보를 추가하였고, [16]은 단어로 구성된 기본 특징에 단어 간의 관계정보를 추가, 확장하여 얻어낸 혼합 특징을 사용하여 성능향상을 기대하였다. 또 관심 지점의 분류에 관한 연구[17]도 있었는데 관심 지점 명칭 단어와 문맥 정보를 활용하여 70% 정도의 정확률을 보여주었다. 스팸 메일 필터링에 관한 연구[18]에서는 메일 문서의 특성인 하이퍼링크 정보가 포함되었다는 점을 이용하여 9.4%의 성능향상을 가져왔다. 그밖에 검색 도구 구축 시 자질로 음소를 사용한 연구[19]도 있었는데 SNS 글이나 댓글처럼 초성으로 표현한 단어들을 포함한 짧은 문서의 분류 등에 쓰일 수 있을 것으로 판단된다. 또 [20]에서 제안한 커널 기반의 구조 자질은 구구조 문법의 언어에서 이용할 수도 있겠다.

중국어 문서분류에 대한 연구로는 단어에 의존관계를 추가하여 좋은 성능을 얻어낸 연구[21]이 있었고, uni-gram을 자질어로 하여 비교적 좋은 결과를 얻어낸 연구[9]도 있었다.

본 연구에서는 기존 연구에서 전반적으로 좋은 성능을 보인 SVM 분류기와 정보획득량 기법을 이용하여 실험을 진행하였다.

중국어 문서분류에서 기본 자질어를 통합적으로 비교한 연구가 없었기 때문에 본 논문에서는 중국어 문서분류에서 어떤 기본 자질어가 가장 좋은 성능을 나타내는지를 알아보려고 한다. 또한 기존의 한국어 기본 자질어에 대한 선행연구와의 비교를 통하여 중국어 문서분류와 한국어 문서분류간의 최적 자질어를 비교하고자 한다.

3. 문서 분류 방법

3.1 실험 데이터

3.1.1 한국어 신문기사

한국어 실험 데이터는 25,392개의 한국어 인터넷 신문기사를 사용하였다. 해당 신문기사들은 정치, 사회, 경제, 스포츠, 연예, 해외연예, 사고, TV/방송 등 9개의 카테고리로 나뉘어 있다. 이 중 신문기사의 수가 600개가 되지 않는 해외연예, 사고와 TV/방송 카테고리의 신문기사들을 제외한 6개 카테고리(총 24,605개)의 신문기사들을 사용하였다. 테스트 데이터로는 각 카테고리당 350개씩 총 2,100개의 신문기사를 사용하였다.

3.1.2 중국어 신문기사

중국어 실험 데이터는 28,836개의 중국어 인터넷 신문기사를 사용하였다. 수집한 신문기사 중에서 멀티카테고리 신문기사 20여개는 제외하였다. 수집한 신문기사의 카테고리는 정치, 경제, 법, 군사, 에너지, 부동산, 스포츠, 연예, 로컬, 세계 등 10개였으나, 로컬과 세계 카테고리의 신문기사는 다른 카테고리와 의미적으로 중복되는 기사가 다수 포함되어 있어 이를 제외하고, 나머지 8개 카테고리의 신문기사 총 20,127개의 신문기사를 사용하였다. 테스트 데이터는 각 카테고리에서 300개씩 추출하여 2,400개의 신문기사를 사용하였다.

3.2 문서분류기 학습

신문기사를 분류하는 SVM 문서분류기를 학습하고 적용하는 절차는 그림 1에 제시하였다.

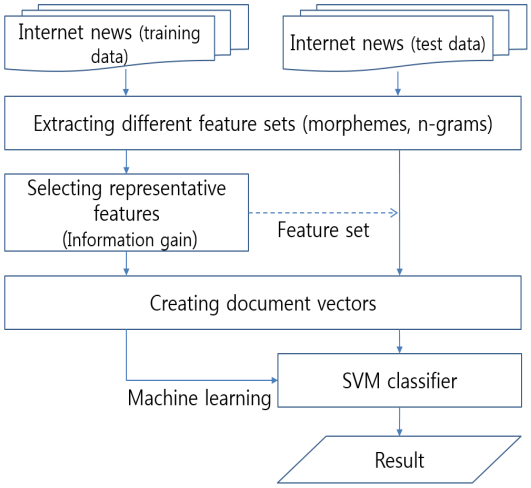


그림 1. 문서분류 절차
Fig. 1. Procedure of Text Classification

신문기사 학습 데이터로부터 형태소와 n-gram과 같은 단위 자질어를 추출하고, 정보획득량 알고리즘을 적용하여 변별력이 높은 자질어 집합을 선정한다. 이 집합을 대상으로 각 문서별 문서벡터를 생성하고, 문서벡터와 각 문서의 카테고리 정보를 가지고 SVM 학습을 하게 된다.

테스트 데이터가 주어지면 학습과 동일한 절차대로 단위 자질어 추출 및 문서 벡터 생성을 하고, SVM 분류를 통하여 최종 결과를 얻게 된다.

한국어 실험에서는 형태소 분석을 수행하여 '명사', '명사·동사·형용사' 집합을 구축하였고, n-gram집합은 bi-gram과 tri-gram집합을 구축하여 테스트를 진행하였다.

중국어 실험에서는 중국어 형태소 분석기의 특수한 결과물인 '사자성어'를 고려하여 '명사', '명사·사자성어', '명사·동사·형용사·사자성어' 집합을 구축하였고, n-gram집합은 uni-gram과 bi-gram집합을 구축하였다. 중국어는 언어의 특성상 띄어쓰기가 없기 때문에 문장의 음절 단위 bi-gram 집합과 형태소 분석 후 형태소 단위 bi-gram으로 나눈 집합 두 가지를 테스트하였다. 예를 들어, "我是学生"(나는 학생입니다.) 라는 구절을 음절 단위 bi-gram으로 나누었을 때는 "我是, 是学, 学生", 이렇게 나뉘지만, 형태소 분석("我/pron. 是/v. 学生/n.") 후 형태소 단위 bi-gram으로 나누면 "我是, 是学生"이 결과물이 된다. 이를 뒤에서 언급하거나 그래프에서 표기할 때에는 'bi_morph'로 표기하기로 한다.

그밖에 복합자질어로 'uni-gram·bi-gram' 자질어와 각 언어의 형태소 자질어 중에서 성능이 가장 좋은 자질어와 n-gram 자질어 중에서 성능이 가장 좋은 자질어를 하나씩 선택하여 새로운 자질어 집합을 구축하여 테스트하였다.

3.3 구현 및 평가척도

한국어 형태소 분석기는 POSTECH KLE 연구실에서 개발한 KOMA를 사용하였고, 중국어 형태소 분석기는 중국 교육부 언어응용 연구실에서 개발한 CorpusWord Parser[22]를 사용하였다. SVM 기계학습을 위한 데이터마이닝 라이브러리는 Waikato 대학의 WEKA[23]을 사용하였다.

평가척도로는 F1-Measure를 사용하였는데, 이는 정확률과 재현율에 동일한 가중치를 부여하여 수식 (1)과 같이 조화평균을 구한 것이다. 10-묶음 교차 검증을 실시하여 실험 결과를 구하였다.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

4. 실험 결과 및 분석

4.1 한국어 자질어별 실험 결과

한국어 신문기사에 대한 자질어 종류별 문서분류 실험 결과는 그림 2에 제시하였다. 단일 자질어 집합인 bi-gram이 가장 좋은 성능을 보여주고 있고 tri-gram이 가장 낮은 성능을 보이는 것으로 나타났다. 명사, '명사·동사·형용사' 등 품사에 따른 자질어 집합은 오분석, 미등록어 문제 등 형태소 분석기의 성능에 영향을 받기 때문에, 일관성 있게 자질어 추출이 가능한 bi-gram의 성능이 상대적으로 좋게 나온 것으로 보인다. tri-gram의 경우는, 학습데이터로부터 추출된 tri-gram이 모든 사례를 포함할 수 없는 데이터부족 현상으로 인해 성능이 낮은 것으로 판단된다.

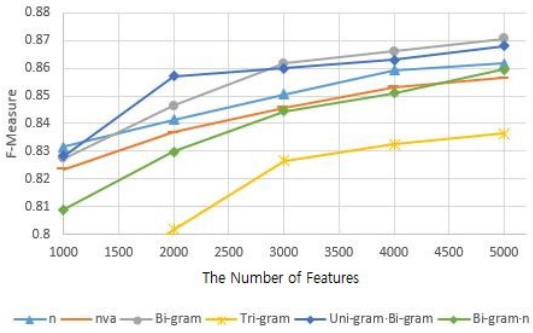


그림 2. 한국어 신문기사 분류 결과
Fig. 2. Results of Korean Text Classification

다른 복합 자질어로는 언어 독립적 자질 가운데 가장 좋은 결과를 보인 bi-gram과 언어 종속적 자질 가운데 가장 좋은 결과를 보인 명사 집합을 통합한 'bi-gram·명사' 집합을 구축하여 실험을 수행해 보았다. 이 집합은 tri-gram 다음으로 낮은 정확도를 보여주어 결과가 썩 좋지 않았는데, 이는 bi-gram 집합과 명사 집합 사이에 동일한 의미를 다양한 형태로 표현하는 중복 자질이 많았기 때문으로 보인다. 본 연

구에서는 정보획득량과 같은 자질어 선정 알고리즘을 통하여 분류기 실행에 적당한 크기의 자질어 집합을 사용하므로 이와 같은 중복 자질이 많은 경우에는 좋지 않은 영향을 준 것으로 판단된다.

만약 'bi-gram·명사' 의 모든 통합 자질을 사용하였다면 성능이 bi-gram과 비슷하거나 다소 좋게 나올 수도 있었겠지만, 현실적으로 너무 큰 자질어 집합을 사용하는 것은 분류기 실행에 있어 메모리 등 하드웨어의 사양이 높을 것을 요구할 뿐만 아니라 시간도 많이 소요되므로 비효율적이다.

4.2 중국어 자질어별 실험 결과

중국어 신문기사에 대한 자질어 종류별 문서분류 실험 결과는 그림 3에 제시하였다. 한국어와 다르게 bi-gram이 가장 낮은 정확도를 나타내고 있는데, 이는 두 언어의 음절 정보량에 차이가 있기 때문이다. 즉 중국어에서 하나의 음절, 한 글자가 제공하는 정보의 양이 한국어에서 하나의 음절이 제공하는 정보의 양보다 크기 때문인데, 이처럼 단위 자질어 정보의 양이 클 경우에는 자질어의 종류도 많아지게 되므로, 이를 모두 추출할 만큼 학습데이터의 양이 충분해야 하지만, 한국어 실험에서의 tri-gram 자질어와 마찬가지로 데이터가 부족했던 것으로 보인다. bi-gram을 제외한 상위 결과를 좀 더 자세히 살펴보기 위해서 종축의 단위를 확대하여 그림 4에 제시하였다.

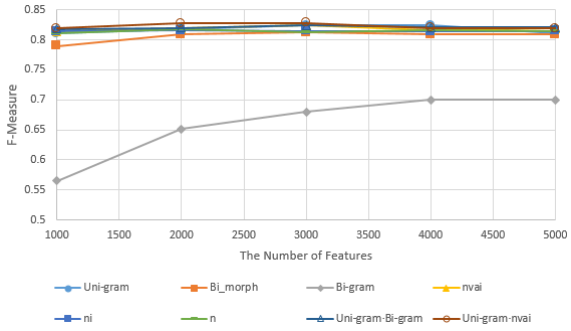


그림 3. 중국어 신문기사 분류 결과
Fig. 3. Results of Chinese Text Classification

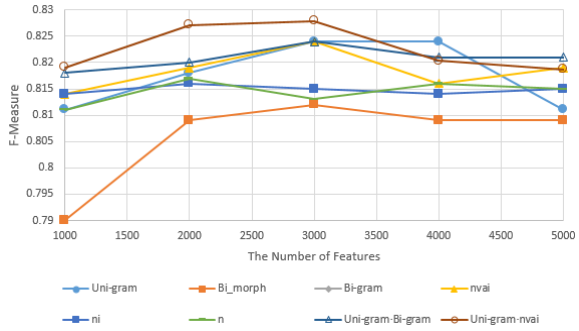


그림 4. 중국어 신문기사 분류 결과 확대
Fig. 4. Detailed View of Chinese Text Classification Results

그림 4를 보면 'uni-gram·명사·동사·형용사·사자성어'와 uni-gram 순으로 좋은 성능을 보여 주고 있다. uni-gram이 좋은 결과에 중복되어 나타나는 것으로 보아, 중국어에서의 uni-gram은 한국어에서의 bi-gram과 비슷한 정보의 양을 가지고 있으며, 중국어 문서분류에 적절한 자질어 단위라고 할 수 있다. 이는 uni-gram이 중국어 자동 문서분류에서 적절한 자질어라는 [9]의 결과와 일치하는 것이다. 중국어 실험에서 명사와 '명사·사자성어' 자질어 집합은 자질어의 개수가 많이 차이나지 않았기 때문에 거의 비슷한 결과를 보여주고 있다. 또 bi_morph 자질어가 bi-gram 자질어보다 좋은 성능을 나타내고 있는 것은 bi_morph 자질어에 형태소 정보가 함축되어 있기 때문으로 보인다.

자질 개수는 2000개~4000개에서 가장 좋은 성능을 보여주고 있는데, uni-gram과 'uni-gram·명사·동사·형용사·사자성어' 집합은 적은 자질수로도 좋은 성능을 보여 주고 있고, '명사·동사·형용사·사자성어' 자질어 집합은 다소 불안정한 결과를 보여준다. 반면 'uni-gram·bi-gram' 집합은 전반적으로 가장 안정된 성능을 보여주고 있다.

4.3 종합 분석

먼저 자질 개수에 따른 성능 변화에 대하여 비교를 하면 중국어는 자질 개수가 2000-4000개 일 때 좋은 성능을 보였고, 한국어는 자질 개수가 5000개 정도일 때 좋은 성능을 보여 주었다. 그 원인을 보면 중국어는 한 글자마다 뜻이 있지만 한국어는 자질 중에 의미어가 아닌 자질(예를 들면 bi-gram에서의 조사 등)이 존재하기 때문에 자질어 집합의 크기가 충분해야 그 경우의 수를 모두 대변하여 성능향상을 기대할 수 있다고 해석 가능하다.

성능이 가장 좋지 않았던 자질어 집합은 두 가지 언어 모두 신문기사 데이터에서 경우의 수가 충분히 표현되지 않아 정보가 부족했던 n-gram이었다. 중국어는 bi-gram이, 한국어는 tri-gram이 각각 성능이 가장 떨어지는 자질어 집합이었다.

이전 절에서 언급한 바와 같이 한국어는 uni-gram, 중국어는 bi-gram의 성능이 전반적으로 좋았고, 또한 [15]에서 서양어권 언어인 영어나 프랑스어가 bi-gram의 성능이 분류결과 정확도 향상에 많이 기여하였다는 점으로 미루어 볼 때, 언어 독립적인 특징인 n-gram이 언어 종속적인 특징인 형태소보다 성능향상에 좀 더 많이 기여한 것으로 여겨진다. 다만 실험 데이터에서 해당 n-gram으로 표현 가능한 특징들이 충분히 표현이 되어야 하므로 n을 정함에 있어서 신중을 기해야 함을 알 수 있다.

언어 종속적인 자질어들을 살펴보면 한국어 실험에서는 명사가, 중국어에서 '명사·동사·형용사·사자성어' 자질어 집합이 가장 좋은 성능을 보여 주었는데 이는 중국어는 개개의 한자마다 뜻을 가지고 있으므로 한국어 음절에 비해 의미 중의성이 존재하고 있기 때문에 문서들을 분류함에 있어서 한국어에 비해 많은 정보를 필요로 하는 것 같다.

두 가지 언어 모두 복합 자질어로 'uni-gram·bi-gram' 자질어 집합을 테스트하였다. 그 결과 두 가지 언어에서 모

두 비교적 높은 정확도를 나타냈다. 특히 한국어에서는 자질어 개수가 적을 때 가장 좋은 성능을 보여주었고 그 뒤로도 가장 좋은 성능을 보여준 bi-gram과 별로 차이나지 않는 결과를 보여주고 있다. 중국어에서는 자질어 개수가 변화함에도 가장 안정적인 성능을 보여주었다. 이는 'uni-gram·bi-gram' 집합이 두 가지 언어에서 모두 적은 자질어로 성능의 최고치에 근접하는 비교적 좋은 결과를 기대할 수 있다는 것을 보여주었다. 또 [15]에 의하면 영문서 분류 결과에서도 'uni-gram·bi-gram' 집합은 해당 연구에서 언급한 네 가지 자질어 집합을 모두 통합한 자질어 집합보다 0.7퍼센트 정도 떨어지는 결과로 두 번째로 성능이 좋았던 점을 감안한다면 비교적 좋은 자질어임을 보여준다. 때문에 'uni-gram·bi-gram' 조합은 여러 언어에서 일반적으로 사용하기에 비교적 적합한 자질어 집합인 것으로 보인다.

또 다른 복합 자질어로 한국어 실험에서는 최적의 언어 독립적 자질인 bi-gram과 최적의 언어 종속적 자질인 명사 집합을 통합한 'bi-gram·명사' 집합을 구축하였고, 중국어 실험에서는 마찬가지로 'uni-gram·명사·동사·형용사·사자성어' 자질어 집합을 구축하여 실험을 하였는데, 한국어 실험에서는 비교적 좋지 않은 결과를 보여주고 있지만, 중국어 실험에서는 이 자질어 집합이 가장 좋은 성능을 보여주고 있다. 이는 한국어는 bi-gram 외에 추가된 형태소정보들이 오히려 그 성능을 떨어뜨리는 작용을 한다는 것을 보여주었다. 하지만 중국어 실험에서는 형태소 정보들이 한자와 그 한자들로 이루어진 단어 간의 의미적 중의성의 해소에 큰 도움이 되었음을 보여준다.

5. 결론 및 향후 연구

본 연구에서는 한국어 및 중국어 신문기사의 문서분류 성능을 높이기 위해 각 언어별 최적의 자질어 단위는 무엇인지 실험을 통하여 살펴보았다. 이를 위해 언어 종속적 자질어인 형태소와 언어 독립적 자질어인 n-gram을 각각 단일 자질어로 추출하여 비교하였고, 또한 언어 종속적 최적 자질어와 언어 독립적 최적 자질어를 조합하여 새로운 복합 자질어 집합을 구축하여 비교하였다.

언어 독립적인 자질어에서는 한국어는 bi-gram이, 중국어는 uni-gram이 가장 좋은 결과를 보여주었고, 언어 종속적인 자질어에서는 한국어는 명사가, 중국어에서 '명사·동사·형용사·사자성어' 자질어 집합이 가장 좋은 성능을 보여 주었다. 특히 중국어에서는 'uni-gram·명사·동사·형용사·사자성어'로 구성된 복합 자질어가 가장 좋은 성능을 보였다.

최적 자질어 집합은 해당 언어로 작성된 문장의 내포 의미를 가장 잘 표현한 형태로 간주할 수 있으므로 문서분류 뿐만 아니라 문서자동요약 등 다른 자연어처리 응용 분야에서도 최적의 자질어로 활용할 수 있겠다.

향후에는 본 연구를 통하여 선정된 최적 자질어를 이용하여 문서자동요약 등 다른 응용서비스에 적용을 할 계획이며, 또한 깊이 있는 자연어처리를 하지 않더라도 간단한 형태의 구문정보와 의미정보를 추출하여 문서의 자질 정보로 추가 활용하는 방법에 대해 연구하고자 한다.

References

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10, pp. 79-86, 2002

[2] B. Kim, "A Study on Comparison with SVM, EM, and Naivebayes Algorithm." *The Institute of Electronics and Information Engineers Summer Conference*, Vol. 32 (1), pp. 683-684, 2009

[3] C. Park, D. Seong, K. Park, "Automatic IPC Classification for Patent Documents using Machine Learning," *Journal of Advanced Information Technology and Convergence*, Vol. 10 (4), pp. 119-128, 2012

[4] X. Li, J. Liu and Z. Shi, "A Chinese Web Page Classifier Based on SVM and Unsupervised Clustering," *Chinese Journal of Computers*, Vol. 24(1), pp. 62-68, 2001

[5] D. Choi, S. Lee, J. Kim, J. Lee, "A Study on Graph-based Topic Extraction form Microblogs," *Journal of The Korean Institute of Intelligent Systems*, Vol. 21(5), pp. 564-568, 2011

[6] T. Basu, C. A. Murty, "Effective Text Classification by a Supervised Feature Selection Approach," I*EEE 12th International Conference on Data Mining Workshops*, pp. 918-925, 2012

[7] Y. Yang and J. O. Pedersen. "A Comparison Study on Feature Selection in Text Categorization," *In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 97)*, pp. 412-420, 1997

[8] B. Shim, J. Park, J. Seo, "Term Weighting Using Date Information and Its Appliance in Automatic Text Classification," *Proceedings of the 19th Annual Conference on Human and Cognitive Language Technology*, Vol. 10, pp. 169-173, 2007

[9] Y. Zhang, J. Lu and J. Yang, "Research on the Technique of Chinese Text Classification Based on the Single Chinese Character Feature," P*attern Recognition*, 2009. CCPR 2009. Chinese Conference on, pp. 1-5, 2009

[10] S. Rho, B. Kim, N. Huh, "Representative keyword

Extraction from Few Documents through Fuzzy Inference," *Journal of The Korean Institute of Intelligent Systems*, Vol. 11(9), pp. 837-843, 2001

[11] T. Goncalves and P. Quaresma, "Text Classification Using Tree Kernels and Linguistic Information,", *IEEE Seventh International Conference on Machine Learning and Applications*, pp. 763-768, 2008

[12] J. Roh, H. Kim, J. Chang, "Improving Hypertext Classification Systems through WordNet-based Feature Abstraction," *Journal of Society for e-Business Studies*, Vol. 18(2), pp. 95-110, 2013

[13] S. Park, B. Zhang, "Text Categorization Using Both Lexical Information and Syntactic Information," *The Korean Institute of Information Scientists and Engineers Autumn Conference*, Vol 28(2), pp. 37-39, 2001

[14] I. Kang, "A Comparative Study on Using SentiWordNet for English Twitter Sentiment Analysis," *Journal of Korean Institute of Intelligent Systems*, Vol. 23 (4), pp. 317-324, 2013

[15] E. D'hondt, S. Verberne, C. Koster and L. Boves, "Text Representation for Patent Classification," *Computational Linguistics*, vol 39(3), pp. 755-775, 2013

[16] J. In, J. Kim, S. Chae, "Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification," *Journal of Korean Society for Internet Information*, vol. 14 (5), pp. 49-57, 2013

[17] S. Choi, S. Park, "Categorization of POIs Using Word and Context information," *Journal of Korean Institute of Intelligent Systems*, Vol 24 (5), pp. 470-476,2014

[18] S. Kang, J. Kim, "Intelligent Spam-mail Filtering Based on Textual Information and Hyperlinks," *Journal of The Korean Institute of Intelligent Systems*, Vol. 14 (7), pp.895-901, 2004

[19] T. Kim, J. Lee, M. Chang, "A Minimal Pair Searching Tool Based on Dictionary," *Journal of The Korean Institute of Intelligent Systems*, Vol. 24(2), pp. 117-122, 2014

[20] J. Son, J. Go, S. Park, K. Kim, "Kernelized Structure Feature for Discriminating Meaningful Table from Decorative Table," *Journal of The Korean Institute of Intelligent Systems*, Vol. 21(5), pp. 618-623, 2011

[21] P. Wang and X. Fan, "Study on Chinese Text Classification Based on Dependency Relation," *Computer Engineering and Applications*, Vol.46(3), pp. 131-141, 2010

[22]H. Xiao, "CorpusWordParser.exe, Computer software. Corpus Online. Vers. 3.0.0.0," *Ministry of Education and Institute of Applied Linguistics*, Web. 〈www.cncorpus.org〉. 2014

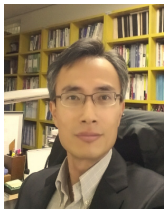
[23] L. H. Witten, E. Frank and M. A. Hall, "DATA MINING: Practical Machine Learning Tools and Techniques," third Edition.

저 자 소 개

임미영(Mei-Ying Ren)

2014년 : 중국 연변대학 과학기술학원 서양어학부 학사
2014년~현재 : 대구대학교 컴퓨터정보학과 석사과정

관심분야 : Natural Language Processing
Phone : +82-53-850-4464
E-mail : meeyeong1211@hotmail.com



강신재(Sinjae Kang)

1995년 : 경북대학교 컴퓨터공학 공학사
1997년 : POSTECH 컴퓨터공학 공학석사
2002년 : POSTECH 컴퓨터공학 공학박사
1997년~1998년 : SK Telecom 정보기술 연구원 주임연구원
2002년~현재 : 대구대학교 컴퓨터-IT공학부 교수

관심분야 : 자연어처리, 온톨로지, 시맨틱 웹
Phone : +82-53-850-6584
E-mail : sjkang@daegu.ac.kr