연구논문

최소거리법과 기계학습법에 의한 한국어 텍스트의 저자 판별*

Author Identification of Korean Texts by Minimum Distance and Machine Learning

金明哲[a] · 허명회[b]
Mingzhe Jin · Myung−Hoe Huh

본 논문은 2개 코퍼스(A, B)의 문자와 기호, 어절, 형태소 태그, 형태소를 단위로 한 n−gram 통계 데이터를 5개의 거리 함수(유클리드 거리, 카이제곱 거리, 가중 유클리드 거리, 코사인 거리, 대칭적 Kullback−Leibler 거리)와 3개의 기계학습법(K−NN, SVM, RF)으로 분석한 한국어 텍스트 저자 판별의 실증적 연구결과를 보고한다. 연구의 결과, SVM(support vector machine)과 RF(random forests)의 판별율이 높았고 코퍼스 A는 최고 98%, 코퍼스 B는 몇 개의 방법이 완벽한 판별율을 기록하였다. 5개 거리 함수 중에서는 가중 유클리드 거리와 대칭적 Kullback−Leibler 거리가 나머지 거리 함수들에 비해 좋은 결과를 보였다.

주제어: 코퍼스의 계량적 분석, 저자 판별, 최소거리법, 기계학습법

This paper reports the author identification (or authorship attribution) study of Korean Corpus A and Corpus B by the quantitative analysis of linguistic characteristics, consisting of letters and symbols, phrases, morpheme tags, and n−grams derived from morphemes. Minimum distance methods(Euclidean, Chi−square, Weighted Euclidean, Cosine, Symmetric Kullback−Leibler) and machine learning methods such as k−nearest neighbor(k−NN), support vector machine (SVM), random forests(RF) are applied to linguistic features extracted from the

* 제1저자는 코퍼스 A 목록을 제공을 중개해 주신 고려대학교 언어학과 최재웅 교수 및 목록을 제공하여 주신 한나래 박사께 감사드린다. 또한 연구넌 환경을 제공해 주신 고려대학교 통계학과의 여러분들께도 감사드린다.

a) 일본 同志社대학(Doshisha University) 문화정보학부 교수

b) 교신저자(corresponding author). 고려대학교 통계학과 교수 허명회.
E-mail: stat420@korea.ac.kr

classified texts. Results show that SVM and RF are superior in recall and precision compared to k−NN and minimum distance methods with the discriminant rate as high as 98% in Corpus A and 100% in Corpus B. Among five distances considered, Weighted Euclidean and Symmetric Kullback−Leibler distances are better than the others.

Key words: quantitative analysis of text corpus, author identification, minimum distance method, machine learning.

Ⅰ. 서 론

우리는 읽은 문장이 소설인지, 논문인지, 신문 기사인지 그 장르를 대충 알 수 있다. 이것은 각 장르의 문장 형식(패턴)을 학습하여 그 지식을 가지고 있기 때문이다. 또한 특정 작가에 관심을 가진 독자는 임의로 주어진 문장이 그 작가의 문장인지 아닌지를 어느 정도 알 수 있다. 그것은 그 작가의 작품을 많이 읽는 동안 작가의 문장 패턴이 독자의 뇌에 기록되기 때문으로 추측된다.

인간 뇌의 패턴 인식 구조는 아직 완전히 해명되지 않았지만, 인간이 하고 있는 패턴 처리와 인식을 컴퓨터로 대체하는 연구가 진행되어 일정한 성과를 얻고 있다. 텍스트에 대한 통계적 패턴 인식도 그 중 하나이다.

텍스트의 저자 판별에 대한 계량적 연구가 시작된 것은 19세기 후반이다. 오하이오 주립대학교의 지구물리학 교수 Mendenhall(1887)은 단어 길이의 분포에 저자의 문체적 특징이 나타난다는 연구결과를 Science지에 발표하였는데, 그는 디킨스(Dickens, 1812~1870), 새커레이(Thackeray, 1811~1863), 밀(Mill, 1806~1873)의 글에 사용된 단어의 길이가 작가에 따라 다르고 이런 것들이 작가의 문체적 특징이 됨을 보였다.

계량적 저자 판별에 관한 동양권의 연구는 일본에서 시작되었다. 계량적 저자 판별에 관한 일본 학자들의 연구는 1950년대에 시작하였고 현재에 이르기까지 다수의 연구결과가 보고되었다. 그 가운데 컴퓨터를 이용한 텍스트의 저자 판별 연구로는 Jin & Murakami(1993), 金(1994), 金(1997), 金(2002), 金・村上(2007)등이 있다.

한국어의 저자 판별에 관한 계량적 연구는 한나래(2009)가 처음이다. 한나래(2009)

는 조선일보에 연재된 칼럼니스트 4인의 텍스트 160개에 대하여 최소거리법으로 저자 판별을 시도하여 최고 93.7%의 판별율을 얻었다. 그는 피어슨 카이제곱 검증 통계량을 사용하였다.

본 연구에서는 카이제곱 거리 외에 네 가지 거리 함수를 추가로 고려하고 세 가지 기계학습법을 활용하여 텍스트 저자 판별을 시도한다. 이 연구결과가 향후 이 분야에서 경험적 지침이 되길 기대한다.

Ⅱ. 텍스트와 특징 데이터

1. 분석 텍스트

본 연구에서는 2개의 코퍼스를 다루었다. 코퍼스 A는 한나래(2009)가 선정한 조선일보 칼럼니스트 4인의 텍스트 모둠이다. 이 코퍼스는 4인의 각 40개 칼럼씩 총 160개 텍스트로 구성되어 있다. <표 1>은 텍스트의 어절 단위 크기에 관한 저자별 요약 통계를 제시한 것이다. 텍스트 제목은 분석에서 제외하였다. <표 1>에서 보듯이 텍스트의 길이에는 큰 차이가 있지 않았다.

코퍼스 B는 박용수 http://blog.naver.com/toamm, 최석영 http://blog.naver.com/6347490, 고연주 http://blog.naver.com/laonella, 한상숙 http://blog.naver.com/sug5205의 2004~2006년 블로그 글 각 40개씩 총 160개 텍스트로 구성된 Web-blog 코퍼스이다. <표 2>는 저자별 텍스트 크기의 요약 통계인데, 네 블로거의 텍스트가 길이에서 큰 차이가 없음을 볼 수 있다. <표 1>과 <표 2>를 비교하면 텍스트의 길이는 코퍼스 B가 코퍼스 A보다 조금 길다.

<표 1> 코퍼스 A의 저자별 텍스트 크기의 요약 통계(어절 단위)

	Mean	Min.	Q1	Median	Q3	Max.
김창규	391.9	368.0	383.0	389.0	399.2	431.0
김대중	476.6	412.0	447.5	461.5	478.5	648.0
류근일	400.8	350.0	390.2	401.0	416.2	441.0
양상훈	423.3	389.0	416.0	422.0	432.0	451.0
전체	423.2	350.0	393.8	417.0	438.2	648.0

〈표 2〉 코퍼스 B의 저자별 텍스트 크기의 요약 통계(어절 단위)

	Mean	Min.	Q1	Median	Q3	Max.
박용수	789.1	505	634.2	717.5	773.2	3358
최석영	712.7	431	568.8	697.0	798.5	1515
고연주	791.0	433	555.0	681.0	989.5	1700
한상숙	601.0	400	509.5	585.0	677.8	944
전체	723.5	400	542.8	664.0	764.0	3358

2. 특징 데이터

텍스트 저자의 통계적 판별에서는 일반적으로 텍스트의 특정 요소를 집계한 데이터 집합을 사용한다. 텍스트에서 특징 데이터 집합을 추출하는 방법은 여러 가지가 가능하다. 한나래(2009)는 (a) 음절, (b) 형태소, (c) 대표 형태소, (d) 비주제 특정 형태소, (e) 품사, (f) 어절, (g) 음절 2연쇄, (h) 형태소 2연쇄, (i) 대표 형태소 2연쇄, (j) 비주제 특정 형태소 2연쇄에 관한 데이터 집합을 분석하여, (b) 형태소, (c) 대표 형태소, (d) 비주제 특정 형태소, (h) 형태소 2연쇄, (i) 대표 형태소 2연쇄의 정답률이 높음을 보고하였다. 한나래(2009)는 한국어 표기 문자 데이터와 품사 2연쇄 데이터는 다루지 않았다.

이러한 점을 감안해, 본 연구는 다음 13개 데이터 집합에 대해 실증적 분석을 했다.

- 문자의 n−연쇄(n−gram), n=1,2,3.
- 어절의 n−연쇄, n=1.
- 품사의 n−연쇄, n=1,2,3.
- 형태소의 n−연쇄, n=1,2,3.
- 비주제 특정 형태소의 n−연쇄, n=1,2,3.

n−연쇄(n−gram)는 기호 열에서 n개 단위의 연쇄 패턴을 일컫는다. n=1인 경우를 unigram, n=2인 경우를 bigram, n=3인 경우를 trigram이라고 한다. 다음 예는 문자가 단위인 n−연쇄를 보여준다.

(예: 데이터세트)

- unigram: 데 이 터 세 트
- bigram : 데이 이터 터세 세트

- trigram : 데이터 이터세 터세트
- four-gram: 데이터세 이터세트

형태소(morpheme)는 의미를 지닌 가장 작은 단위이다. 형태소 분석을 컴퓨터로 할 수 있는데, 이때 쓰이는 소프트웨어가 형태소 분석기(morphological analyzer)이다. 본 연구에서 사용된 형태소 분석기는 POSTAG/Sejong for Windows (http://isoft.postech.ac.kr/Course/CS730b/2005/index.html)이다. 한나래(2009)의 연구에 사용된 분석기와는 다르다.

대표형 형태소라 함은 문법적 형태소들의 異형태(예: "었", "았", "ㅆ")를 대표형(예: "었")으로 통합한 것을 일컫는다(한나래 2009). 본 연구에서는 대표형 형태소는 다루지 않았다. 그 이유는 저자의 습관에 따라 사용하는 형태가 다를 수 있고 때로는 그 다름이 저자 판별에 중요한 정보이기 때문이다. 그리고 대표형 형태소를 이용한 저자 판별율이 비주제 특정 형태소를 이용한 저자 판별율과 유사하기 때문이다(한나래 2009).

비주제 특정 형태소는 모든 형태소에서 품사에 의거해 동사, 일반명사, 고유명사, 외국어, 한자어를 제외한 형태소이다(한나래 2009). 본 연구에서 동사는 제외되지 않았다. 동사는 명사만큼 텍스트의 내용에 의존하지 않으며, 실험 결과 동사를 포함하는 것이 저자 판별에 도움이 되기 때문이다.

텍스트 i에서 집계한 특징에 관한 패턴의 빈도 벡터 $(f_{i1}, \cdots, f_{ij}, \cdots, f_{iM})$은 텍스트 길이에 의존한다. 따라서 본 연구에서는 텍스트 길이에 의존하지 않도록 상대빈도 $(z_{i1}, \cdots, z_{ij}, \cdots, z_{iM})$, 즉 프로파일 벡터로 변환된 텍스트 특징 데이터를 분석하였다. 여기서

$$z_{ij} = f_{ij} / \sum_{j=1}^{M} f_{ij}, \quad j = 1, \cdots, M \tag{1}$$

이다.

이상의 방식으로 구성된 특징 데이터의 크기는 〈표 3〉과 같다. 코퍼스별로 총 160개 텍스트에 걸쳐 출현 수가 10 이하인 것은 '기타(others)'로 처리하였다. 그런 이유로 같은 종류의 특징 데이터라도 코퍼스에 따라 특징(=변수)수가 다르다.

〈표 3〉 코퍼스 · 특징 데이터별 변수수

		코퍼스 A	코퍼스 B
문자와 기호	unigram	802	1,008
	bigram	4,301	6,247
	trigram	2,244	3,707
어절	unigram	860	1,424
형태소 태그	unigram	44	44
	bigram	434	483
	trigram	1,645	1,746
형태소	unigram	1,519	2,095
	bigram	1,947	3,207
	trigram	927	1,753
비주제 형태소	unigram	1,124	1,531
	bigram	339	720
	trigram	72	78

Ⅲ. 저자 판별의 통계적 방법

계량적 저자 판별에서는 텍스트에서 저자의 특징 데이터 집합을 만들고 데이터 집합에 통계적 방법을 적용하여 저자를 판별한다. 저자의 특징을 포함한 데이터 집합을 어떻게 구성하는가도 중요하지만 어느 방법을 쓰는가도 동등하게 중요하다.

통계적으로 텍스트 저자를 분석하는 방법은 비지도학습(unsupervised learning)법과 지도학습(supervised learning)법으로 나누어질 수 있다. 비지도학습에서는 텍스트 저자의 정보 없이 자료 탐색으로 저자를 판별하고, 지도학습에서는 텍스트 저자의 정보가 포함된 연습자료(training data)를 분석하여 저자를 판별한다.

본 연구에서는 주로 지도학습의 기계학습법에 의한 저자 판별법을 활용하였다. 그런데 선행연구인 한나래(2009)에서 최소거리(minimum distance)법이 쓰였으므로, 비교를 위해 그 방법을 본 연구에 포함시켰다.

1. 최소거리법

저자 k 텍스트의 특징 데이터 중심을 $o_k = (o_{k1}, o_{k2}, \cdots, o_{kM})$, 판별대상 텍스트의 중심을 $x = (x_1, x_2, \cdots, x_M)$이라고 하고 두 프로파일 벡터 간 거리를 $d(x, o_k)$로 할 때 $d(x, o_k)$가 최소인 저자 k를 찾는 방법이 최소거리법이다.

두 프로파일 벡터 간 거리로서 널리 알려진 것은 다음과 같이 정의되는 유클리드 거리(ED: Euclidean Distance)이다.

$$d_{ED}(x, o_k) = \sqrt{\sum_{j=1}^{M}(x_j - o_{kj})^2} \quad (2)$$

가중 유클리드 거리(WED: Weighted Euclidean Distance)는 유클리드 거리의 확장이다.

$$d_{WED}(x, o_k) = \sqrt{\sum_{j=1}^{M} w_j(x_j - o_{kj})^2} \quad (3)$$

정규화 유클리드 거리도 그 중 하나이다. 또한 마할라노비스거리(Mahalanobis distance)도 생각할 수 있다. 한나래(2009)가 사용한 카이제곱 거리

$$d_{\mathrm{Chi}D}(x, o_k) = \sqrt{\sum_{j=1}^{M}(x_j - o_{kj})^2 / o_{kj}} \quad (4)$$

는 일종의 가중 유클리드 거리인데, 이것이 산출되려면 기대빈도 o_{kj}가 영이 아니어야 하므로 일정 값 $\alpha(>0)$를 더할 필요가 있다.[1] 본 연구에서는 카이제곱 거리대신 다음의 가중 유클리드 거리를 사용한다.

$$d_{WED}(x, o_k) = \sqrt{\sum_{j=1}^{M}(x_j - o_{kj})^2 / (x_j + o_{kj})} \quad (5)$$

---

1) 한나래(2009)는 상대빈도 셀에 1을 더하여 카이제곱 거리를 산출하였는데 이보다는 관측빈도 셀에 1을 더하여 거리를 산출하는 것이 바른 방법이다.

패턴 인식 분야에서는 코사인 거리도 많이 사용된다. 코사인 거리 함수는 다음과 같다.

$$d_{CosD}(x, o_k) = 1 - \frac{\sum_{j=1}^{M} x_j o_{kj}}{\sqrt{\sum_{j=1}^{M} x_j^2 \sum_{j=1}^{M} o_{kj}^2}} \quad (6)$$

정보학에서는 확률분포 간 거리로 KL(Kullback−Leibler) 정보량이 빈번하게 사용된다.

$$d_{KL}(x, o_k) = \sum_{j=1}^{M} x_j \log_e \frac{x_j}{o_{kj}} \quad (7)$$

KL 정보량은 비대칭적이므로 엄밀하게는 일반적 거리라고 할 수 없다. 그래서 KL 정보량을 대칭이 되도록 재정의한 다음 식을 텍스트의 프로파일 벡터 간 거리로 정의하였다.

$$d_{KLD}(x, o_k) = \frac{1}{2} \sum_{j=1}^{M} \left(x_j \log_e \frac{2x_j}{x_i + o_{kj}} + o_{kj} \log_e \frac{2o_{kj}}{x_i + o_{kj}} \right) \quad (8)$$

여기서 $x_j = 0$이면 $\log_e \frac{2x_j}{x_i + o_{kj}} = 0$으로 하고 $o_{kj} = 0$이면 $\log_e \frac{2o_{kj}}{x_i + o_{kj}} = 0$으로 한다.

본 연구에서는 유클리드 거리 d_{ED}, 카이제곱 거리 d_{ChiD}, 가중 유클리드 거리 d_{WED}, 코사인 거리 d_{CosD}, 대칭적 Kullback−Leibler 거리 d_{KLD}를 사용한다.

2. 기계학습법

기계학습(machine learning)은 인간의 학습 능력을 컴퓨터에서 실현하고자 하는 공학적 기법이다. 데이터 집합을 분석하여 실용적 규칙과 판단 기준 등을 만들어낸다. 통계학이 다루는 판별분석과 회귀분석도 일종의 기계학습으로 볼 수 있다. 그러나 본고에서는 기계학습을 데이터 마이닝과 인공지능 등에서 개발된 공학적 방법을 지칭하는 용어로 사용한다.

앞에서 언급했듯이 기계학습에는 비지도학습과 지도학습이 있으며, 지도학습에도 많은 방법이 있다. 텍스트의 분류에 관해서는 Sebastiani(2002)가 주요 연구결과를 총괄한 바 있다. 지도학습의 분류 정확도는 텍스트 종류와 특징 데이터의 구성 방식 등에 의존하므로 절대적인 것은 아니지만, 일반적으로 k−NN(k−nearest neighbor), SVM (support vector machine)과 앙상블 학습(ensemble learning) 등이 좋은 것으로 알려져 있다. 일본어 텍스트의 저자 판별에 관해서 金・村上(2007)는 몇 가지 지도학습법을 적용한 결과 SVM과 앙상블 학습이 우수하다고 보고한 바 있다.

1) k-근접 이웃

k−근접 이웃(k−NN)의 알고리즘은 여러 기계학습 중에서 가장 간단하다. k−NN은 판별대상 텍스트와 가장 가까운 k개의 학습 텍스트(케이스) 간 거리를 계산한 후, k개 케이스들의 투표로 판별대상 텍스트를 분류한다. 최적 k는 코퍼스의 종류와 학습 텍스트의 수 등에 따라 다르다. 본 연구에서는 유클리드 거리를 사용하였고 탐색적 실험을 통해 k를 7로 정했다.[2)]

2) 서포트 벡터 머신

서포트 벡터 머신(SVM)은 Vapnik(1995)이 고안한 패턴 분류 방법이다. SVM은 학습 개체들을 그룹별로 분리하는 초평면 경계 간 폭, 즉 마진(margin)을 기준으로 최적의 분류규칙을 찾는다. 본 연구에서는 커널 SVM을 썼고 커널 함수는 RBF(radial basis function)로 하였다.[3)]

3) 앙상블 학습

앙상블 학습이란, 다수의 기본 학습기를 생성하여 그것들을 결합하여 새로운 학습기를 만드는 방법이다. 앙상블 학습기는 기본 학습기의 예측을 비가중 투표나 가중치 결합으로 종합한다. 앙상블 학습의 대표적인 방법으로 부스팅(boosting), 배깅(bagging), Breiman (2001)의 랜덤 포리스트(RF: random forest) 등이 있다. RF는 특수한 경우를 제외하면 배깅과 부스팅에 비교하여 같거나 우수한 결과를 내는 것으로 알려져 있다. 본 연구에서는 RF 앙상블 학습을 적용하였는데 이 RF는 전체 변수 수

---

2) R 2.15.1에서 class 패키지의 knn 함수를 활용하였다.

3) R 2.15.1에서 kernlab 패키지의 ksvm 함수를 활용하였다. 파라미터 sigma는 디폴트값으로, 파라미터 C는 1로 두었다.

의 제곱근만큼의 변수들로 생성된 1,000개의 기본 분류나무로 만들어졌다. 전체 데이터의 63%가 연습(training)에 할당되었고 나머지 37%가 테스트에 할당되었다.[4)]

3. 판별 성능 평가

코퍼스가 G개 그룹의 텍스트로 구성되어 있다고 하자. 이 연구에서 G는 저자 수와 일치한다. 그룹 $g\,(=1, \cdots, G)$의 텍스트들에 대한 판별 결과는 〈표 4〉와 같은 혼동 행렬(confusion matrix)로 요약될 수 있다.

그룹 g의 판별 결과는 재현율(recall) 및 정확률(precision) 등으로 평가된다. 재현율은 판별 결과가 "알짜배기"의 누출 없이 얼마나 알짜배기를 올바르게 선별하고 있는지에 대한 정도이고 정확률은 "알짜배기"를 고른 결과에 얼마나 알짜배기가 포함되는지에 대한 정도이다. 그룹 g의 재현율과 정확률의 수식 정의는 다음과 같다.

$$\text{그룹 } g \text{ 재현율: } R_g = \frac{a_g}{a_g + c_g}\,, \text{ 그룹 } g \text{ 정확률: } P_g = \frac{a_g}{a_g + b_g} \tag{9}$$

본 연구에서 판별 결과의 총괄적 평가는 개별 그룹 판별의 재현율과 정확률의 매크로 평균(macro average)으로 하였다. 매크로 평균의 수식 정의는 다음과 같다.

$$\text{총 재현율: } \overline{R} = \frac{1}{G}\sum_{g=1}^{G}\frac{a_g}{a_g + c_g}\,, \text{ 총 정확률: } \overline{P} = \frac{1}{G}\sum_{g=1}^{G}\frac{a_g}{a_g + b_g} \tag{10}$$

〈표 4〉 그룹 g에 대한 판별 결과의 혼동 행렬

그룹 g 인가?		판별	
		예	아니오
실제	예	a_g	c_g
	아니오	b_g	d_g

4) R 2.15.1에서 randomForest 팩키지의 randomForest 함수를 활용하였다.

본 연구에서는 판별 결과의 평가는 총 재현율과 총 정확률의 조화평균 F_1을 퍼센트로 표현하여 쓴다.

$$F_1 = \frac{100}{\frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right)} \tag{11}$$

이 연구에서 이를 판별율로 칭한다. 따라서 F_1 값이 큰 판별기가 좋은 평가를 받게 된다.

판별 성능은 개별 코퍼스 160개에 교차검증(leave−one−out cross−validation)법을 적용하여 평가하였다.

Ⅳ. 판별 결과

1. 코퍼스 A

코퍼스 A에서 산출된 F_1 값을 〈표 5〉에 정리하였다. 판별방법 간 차이의 유의성에 대해 고찰하기 위해 방법들 간 평균차이에 관한 양측 t 검정을 실시하여 〈표 6〉에 제시하였다.

최소거리법에서는 ChiD와 WED 및 KLD 사이에 유의한 차이가 있었다($\alpha = 5\%$). 그러나 WED와 KLD 간 차이는 유의하지 않았다.

기계학습법인 SVM와 RF 사이에도 유의한 차이가 없었다. 그러므로 최소거리법에서는 WED와 KLD가, 기계학습법에서는 SVM과 RF가 코퍼스 A에 최적이라고 판단한다. 따라서 본고의 이하 특징 데이터 분석은 WED, KLD, SVM, RF의 F_1 값을 중심으로 한다.

문자기호의 n−gram에서는 전체적으로 보면 bigram의 F_1 값이 크다고 할 수 있다. 그러나 SVM과 RF의 경우, unigram과 bigram 간에는 큰 차이가 없었다. bigram의 SVM 판별율이 가장 높은 F_1 값을 보였다(96.30%).

〈표 5〉 코퍼스 A의 F_1 값

방법		최소거리					기계학습		
		ED	ChiD	WED	CosD	KLD	KNN	SVM	RF
문자와 기호	unigram	88.01	86.78	90.62	87.43	89.41	60.89	95.87	96.25
	bigram	55.61	87.51	93.16	86.19	93.16	77.69	**96.30**	96.29
	trigram	48.36	77.94	91.21	77.39	91.21	64.74	95.70	93.79
어절	unigram	47.03	82.63	87.99	80.43	86.90	69.18	**97.54**	96.27
형태소 태그	unigram	88.26	89.37	91.35	87.04	90.74	86.59	92.51	92.49
	bigram	86.18	91.30	95.05	86.90	93.86	86.64	97.51	96.27
	trigram	81.25	92.67	95.21	83.67	94.54	71.16	96.26	96.31
형태소	unigram	78.22	90.62	92.58	85.65	91.41	82.7	95.65	97.55
	bigram	53.02	85.72	95.03	80.68	95.03	77.34	**98.13**	96.26
	trigram	44.78	69.30	84.39	65.85	86.41	65.88	95.01	91.32
비주제 형태소	unigram	89.46	92.51	**96.90**	90.09	95.66	82.51	94.60	96.23
	bigram	66.66	90.66	94.69	81.27	95.80	76.25	96.30	93.77
	trigram	46.98	77.07	91.42	69.15	90.64	60.98	95.12	89.83

〈표 6〉 판별방법별 F_1 값 평균과 t 검정의 p-값: 코퍼스 A의 경우

	평균	ED	ChiD	WED	CosD	KLD	k-NN	SVM
ED	67.22							
ChiD	85.70	0.0040						
WED	92.28	0.0003	0.0078					
CosD	81.67	0.0182	0.1653	0.0002				
KLD	91.91	0.0004	0.0106	0.7737	0.0002			
k-NN	74.04	0.2475	0.0015	0.0000	0.0280	0.0000		
SVM	95.88	0.0001	0.0002	0.0028	0.0000	0.0006		
RF	94.82	0.0002	0.0006	0.0376	0.0000	0.0134	0.0000	0.1807

어절에서는 SVM의 F_1 값이 제일 높게 나타났다(97.54%). 한나래(2009)는 재현율로 판별 결과를 평가하였고, 어절 데이터의 재현율은 82.5%였다. 본 연구가 얻은 카이제곱 거리(ChiD)의 재현율은 82.50%, 정확률은 82.76%이므로 재현율은 한나래의 결과와 같다.[5] 카이제곱 거리(ChiD)에 비교하여, 가중 유클리드 거리(WED)가 5%p 정도, KLD 거리가 4%p 정도 성능이 좋은 것으로 나타났다.

형태소 태그에서는 bigram의 F_1 값과 trigram의 F_1 값이 unigram의 F_1 값보다 컸다. bigram의 F_1 값과 trigram의 F_1 값 간에는 큰 차이가 없었으나 bigram의 SVM에서 F_1 값이 가장 크게 나타났다(97.51%).

형태소에서는 bigram의 SVM의 F_1 값이 가장 컸고(98.13%), 이 값은 전체에서 가장 크다. RF에서는 bigram이 unigram보다 조금 낮았다.

비주제 형태소의 F_1 값은 판별 방법에 따라 조금 다른 경향을 보였다. 기계학습법에서는 비주제 형태소의 F_1 값이 형태소의 값에 비해 조금 떨어지지만, 최소거리법에서 unigram을 보면 비주제 형태소의 값이 형태소의 값에 비해 4%p 정도 높았다. 이것은 노이즈의 영향에 약한 거리법의 특성 또는 코퍼스 A의 특징이 원인일 가능성이 있다.

2. 코퍼스 B

코퍼스 B에서 산출된 F_1 값을 <표 7>에 정리하였다. 판별방법 간 차이의 유의성에 대해 고찰하기 위해 방법들 간 평균차이에 관한 양측 t 검정을 실시하여 <표 8>에 제시하였다. 그 결과, 코퍼스 A에서와 비슷한 경향을 볼 수 있었다. 즉, 최소거리법에서는 WED와 KLD가 가장 좋았고 두 방법 간 차이는 유의하지 않았다. 그리고 SVM와 RF 사이에도 유의한 차이가 없었다.

문자와 기호에서 최소거리법은 unigram의 F_1 값이 높았고, 기계학습법에서는 trigram의 F_1 값이 높았다. 어절에서는 RF의 F_1 값이 높았다. 형태소 태그에서 WED와 KLD는 bigram의 F_1 값이, SVM와 RF는 trigram의 F_1 값이 높았다. 형태소에서는 unigram의 경우가 전체적으로 높았다. 그러나 기계학습법에서는 bigram과 큰 차이가 없었다. 비주제 형태소에서도 unigram의 경우가 전체적으로 높았다.

---

5) 본 연구와 한나래(2009)의 연구 간 비교가능한 항목은 어절 데이터뿐이다.

〈표 7〉 코퍼스 B의 F_1 값

방법		최소거리					기계학습		
		ED	ChiD	WED	CosD	KLD	KNN	SVM	RF
문자와 기호	unigram	93.84	95.68	99.38	93.83	100.0	77.85	92.81	100.0
	bigram	66.76	91.08	98.77	89.58	98.77	54.99	99.38	99.38
	trigram	56.18	88.15	91.24	86.09	92.46	64.18	100.0	100.0
어절	unigram	50.41	93.81	94.62	91.22	92.47	76.94	98.78	100.0
형태소 태그	unigram	88.26	89.37	91.35	87.04	90.74	87.05	91.32	92.49
	bigram	88.24	91.29	94.03	88.31	92.91	78.45	98.15	98.12
	trigram	79.74	90.79	94.63	85.81	94.59	76.00	96.95	97.53
형태소	unigram	82.59	95.88	98.16	89.98	98.16	74.93	100.0	99.38
	bigram	60.05	87.15	91.89	82.37	92.53	65.30	99.38	99.38
	trigram	50.91	77.12	85.76	70.72	86.76	64.18	95.00	96.30
비주제 형태소	unigram	92.60	96.32	96.97	91.98	96.97	75.36	100.0	100.0
	bigram	64.30	88.87	94.62	82.02	95.18	50.89	98.76	98.77
	trigram	45.05	77.57	88.62	70.01	89.16	69.52	96.26	91.90

〈표 8〉 판별방법 별 F_1 값 평균과 t 검정의 p-값: 코퍼스 B의 경우

	평균	ED	ChiD	WED	CosD	KLD	k-NN	SVM
ED	70.69							
ChiD	89.47	0.0024						
WED	93.85	0.0004	0.0438					
CosD	85.31	0.0137	0.1348	0.0019				
KLD	93.90	0.0004	0.0399	0.9742	0.0017			
k-NN	70.43	0.9646	0.0000	0.0000	0.0003	0.0000		
SVM	97.45	0.0001	0.0006	0.0155	0.0000	0.0145	0.0000	
RF	97.94	0.0001	0.0003	0.0065	0.0000	0.0059	0.0000	0.6571

Ⅴ. 결 언

본 논문에서는 코퍼스 A(조선일보 칼럼)과 코퍼스 B(웹 블로그)에 5개의 거리 함수와 3개의 기계학습법을 적용하여 한국어 텍스트의 저자 판별에 대한 실증 분석을 하였다. 연구의 결과, SVM와 RF 기계학습법이 최소거리법보다 판별율이 높았고, 코퍼스 A는 최고 98%, 코퍼스 B는 몇 개의 방법이 완벽한 판별율을 보였다. 코퍼스 A가 코퍼스 B보다 판별율이 낮은 것은 텍스트 패턴이 상당히 정해져 있는 신문에 게재하는 칼럼니스트 텍스트라는 장르적 특성이 원인일 것으로 추측한다.

본 연구에서 얻은 판별율은 절대적인 것은 아니다. 텍스트 특징 벡터의 구성방식과 기계학습의 파라미터의 조정을 통해 더 높은 판별율을 얻을 가능성이 있다.

5개의 거리 함수에서는 가중 유클리드 거리 WED와 대칭적 Kullback－Leibler 거리 KLD가 다른 거리 함수보다 판별율이 높았다. 이 거리들은 한나래(2009)의 카이제곱 거리와 유의한 차이를 보였다.

특징 데이터 중 어떤 데이터를 사용하는가에 관해서는 이용하는 판별법에 따라 다르기에 결론을 내리기가 어렵다. 그리고 판별율이 가장 높게 나왔다고 그 방법만을 사용하는 것은 옳지 않다. 그렇게 단순한 문제가 아니다. 어떤 특징 데이터에서도 오(誤)판별의 가능성이 있기 때문에 다양한 관점에서 고찰하는 것이 필수적이다. 특히 특징 데이터 사이에 언어학이나 심리학 등의 측면에서 상관(相關)이 적은 다양한 특징 데이터를 종합적으로 활용할 필요가 있다.

한국어 텍스트의 저자 판별에 관한 연구는 초기 단계이고 따라서 많은 과제가 남아 있다. 향후 과제로는 한국어의 고유한 특성을 살린 저자의 특징 데이터 추출 방법의 모색, 텍스트의 길이와 판별율과의 관계, 학습 텍스트 수와 판별율과의 관계, 특징 데이터에 적합한 판별 방법 개발, 여러 특징 데이터와 여러 판별법에 의한 종합적 판별 방법의 개발 등이 있다.

참고문헌

Breiman, L. 2001. “Random Forests.” *Machine Learning*. 45(1): 5−32.

Jin, M and M. Murakami. 1993. “Author's Characteristic Writing Styles as Seen Through Their Use of Commas.” *Behaviormetrika* 20(1): 63−76.

Mendenhall, T.C. 1887. “The Characteristics Curves of Composition.” *Science* IX: 237−249.

Sebastiani, F. 2002. “Machine Learning in Automated Text Categorisation.” *ACM Computing Surveys* 34(1): 1−47.

Vapnik, V.N. 1995. *The Nature of Statistical Learning Theory*. N.Y.: Springer−Verlag.

金明哲. 2002. “助詞のn−gramモデルに基づいた書き手の識別.” ≪計量国語学≫ 23(5): 225−240.

金明哲. 1994. “読点の打ち方と文章の分類.” ≪計量国語学≫ 19(7): 317−330.

金明哲. 1997. “助詞分布に基づいた日記の書き手の認識.” ≪計量国語学≫ 20(8): 357−367.

金明哲・村上征勝. 2007. “ランダムフォレスト法による文章の書き手の同.” ≪数理統計≫ 55(2): 255−268.

한나래. 2009. “빈도 정보를 이용한 한국어 저자 판별.” ≪인지과학≫ 20(2): 225−241.

<접수 2012/9/22; 수정 2012/11/5; 게재확정 2012/11/12>