http://dx.doi.org/10.7236/JIIBC.2013.13.6.155
JIIBC 2013-6-20

나이브 베이지안을 사용한 성명에 대한 성별 구분 연구

A Study on Sex Classification of a Name using Naive Bayesian

임명재*, 정진표**, 김명관***

Myung-Jae Lim, Jin-Pyo Jung, Myung-Gwan Kim

요 약 본 논문은 Naive Bayesian분류기를 사용하여 성명의 성별을 구분하는 시스템을 구현 하였다. 국내인 성명은 외국인 성명과는 다르게 사람을 지칭할 때 쓰는 대명사의 성별불일치 현상이 있다. 하지만 국내인 성명의 특성으로 남자로 자주 쓰이는 이름과 여자로 자주쓰이는 이름을 구분하게 하였다. 그리고 고유명사등, 성별이 애매한 이름들도 포함하였기 때문에 다소 정확율이 떨어지는 것을 확인 할 수가 있었다. 본 논문의 실험 결과로는 국내인 남자는 84%, 여자는 88%의 정확율을 보였으며, 총합 86%의 정확율과 외국인 성명은 남자는 80%, 여자는 84%로 총합 83%의 정확율을 보이고 있다.

Abstract This article employs Naive Bayesian Classifier to realize a system that can distinguish the sex of a name. Unlike foreign names, in Korean names, the pronoun referring to a person shows discordance with sex. With the characteristics of Korean names, however, the study distinguishes names frequently used for men and for women. And as it also includes names of which sex is rather ambiguous such as proper nouns, the accuracy of it is somewhat low. The result of the experiment conducted in this article indicates 84% accuracy for Korean men and 88% for Korean women; thus, the total accuracy equals 86%. Meanwhile, about foreign names, men show 80% accuracy, and women 84%, so the total accuracy equals 83%.

Key Words : Naive Bayesian, Person's names, Classification, Gender

Ⅰ. 서 론

질의-응답시스템은 문서로부터 사용자가 원하는 해답을 찾아 제공해주는 시스템이다. 질의 중에는 사람의 이름을 묻는 질의, 회사나 기관과 같은 조직 이름을 묻는 질의, 지명을 묻는 질의, 시간이나 거리와 같은 단위를 묻는 질의 등이 있다.[2] 문서로부터 해답을 추출하기 위해서는 사람이름, 조직 이름, 지명, 단위 등과 같은 개체명을 인식[3]하고 추출하는 방법이 필요하다. 질의 중에는 '<영화 제목>의 여자 주인공은?' 과 같이 남녀 이름을 구분하여 묻는 경우가 있다. 따라서 사람 이름을 남녀로 구분하여 인식할 필요가 있다. 본 논문에서는 한국 사람의 이름 특성을 이용하여 사람 이름을 인식하고 더 나아가 한국 성씨를 가진 외국인들까지 이름을 인식하여 성별을 구분하여 인식하는 방법에 대해 제안한다. 사람 이름 인식과 성별 구분을 위해 성씨와 이름에 나타나는 음절의 특성 및 통계 정보[4]를 이용한다.

본 논문에서는 국내인 성명과 외국인 성명을 학습 하

*종신회원, 을지대학교 의료IT마케팅학과
**준회원, 을지대학교 의료IT마케팅학과
***정회원, 을지대학교 의료IT마케팅학과
접수일자 :2013년 11월 8일, 수정완료 : 2013년 12월 8일
게재확정일자 : 2013년 12월 13일

Received: 8 November, 2013 / Revised: 8 December, 2013
Accepted: 13 December, 2013
***Corresponding Author: binsum@eulji.ac.kr
Dept. of Medical IT & Marketing, Eulji University, Korea

고 질의 응답 시스템을 사용하여 이름을 인식하고, 성별을 구별하는 실험을 한다. 실험에는 Naive Bayesian 분류기를 사용하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 성명 인식에 대한 기존 연구에 대해 알아본다. Ⅲ장에서는 성명의 특성에 대해 알아본다. Ⅳ장에서는 성명에 대한 성별 분류에 대해 설명하였다. Ⅴ장에서는 사람 이름 인식과 성별 구분 실험을 수행하고, 실험 결과를 분석한다. Ⅵ장에서는 결론 및 향후 연구에 대해 토론하였다.

Ⅱ. 성명 인식에 대한 기존 연구

기존 국내 연구 관련 논문은 신문, 잡지 등 여러 문서들의 특정이름에 대한 성별추론에 관한 연구로 영어권 내의 문서에는 he, she와 같은 성별 표시[3]가 되지만 국내 신문 등에는 직업명사에 '여의사' 혹은 '여교수'등의 표지를 붙이는 경우가 많았으나 근래에는 드물게 사용되고 있다고 한다.[6] 신문기사의 경우 남성과 여성을 '그'로 통일 시키거나 대명사 사용을 기피하고 직함을 반복 사용함으로써 별도의 첨부된 사진이 없는 한 성별 추론은 고유명사[6]인 이름과 직업명사에 대한 개인의 성별 고정관념에 의지하여 이뤄질 수밖에 없다. 그러나 이들에 대한 고정관념이 맥락의 일관성 유지를 위해 지속적이고 강하게 작용할 경우 지시대상의 성별은 실제와 무관하게 통속적인 고정관념에 의해 쉽게 처리되거나 독자의 개인적인 해석에 의존하여 객관성을 읽게 될 가능성이 제기되어 성별표지가 없고 직업명사가 빈번하게 출연하는 한국 문서들에서 한국어 이름의 성별정보가 성별표지의 절대적인 기준이 되는지를 검증하고자 하는 내용이다.[7]

그리고 국내인 성명의 성씨와 이름에 나타나는 음절의 통계 정보를 이용하여 성씨 사전 상위 100개 미만의 성씨만을 이용하여 국내인 성명의 두 번째와 세 번째에 나타나는 음절의 종류와 각 빈도수를 계산하고 남녀 이름에 나타나는 특성을 이용하여 데이터를 구축 테스트 하였다.[2]

그림 1은 휴모션에서 개발한 젠더모션은 2008년도에 V1.0을 공개하였다. 젠더모션(GenderMotion V1.0)은 성명을 입력하면 남자, 여자를 구분하여 주지만 외국인 성명에 대한 구분은 정확도가 많이 떨어지고 있다. 그리고 2음절 이하의 성명을 입력할 시에 인식이 불가능하였고 향후 업데이트가 안 되고 있다.[9]

그림 1. 젠더모션
Fig. 1. GenderMotion

그림 2는 이루미 작명의 홈페이지이다. 이름을 입력시 남녀가 어느 정도의 비율로 이름을 사용하고 있는지, 자신의 이름이 가장 많은 사용되고 있는 성씨가 어느 성씨 인지 등의 통계를 보여주고 비슷한 이름까지 알려준다.[9]

그림 2. 이루미작명 이름통계
Fig. 2. Statistics for Person's name

Ⅲ. 성명의 특성

국내인 성명은 100대 성씨가 차지하는 비율은 전체의 99.1%이고 96%이상이 3음절로 이루어져있고 성씨와 이름으로 구분되며[2], 외국인의 이름은 first name, middle name, last name(family name)으로 나눠진다. 그리고 middle name은 선택적으로 사용되고 있다. 표 1[10]은 미국의 성씨 분포도를 보여준다. 한국과 다르게 쉽게 성을 만들어 쓸 수 있기 때문에 그 숫자는 600만

여 가지의 성씨가 존재한다.

표 1. 미국 성씨 분포도
Table 1. Statistics for American Names

미국 성씨	
Frequency of Occurrence	Number
1,000,000+	7
100,000~999,999	268
10,000~99,999	3,012
1,000~9,999	20,369
100~999	128,015
50~99	105,609
25~49	166,059
10~24	331,518
5~9	395,600
2~4	1,056,992
1	4,040,966

국내인 성명의 이름은 성씨로 사용되는 음절의 수가 제한적이고 2000년 통계청 인구조사 기준 286개이고 표 2 는 한국 성씨 분포도를 보여준다. 국내인 성명에는 한자 독음이 많이 사용되며, 남자 이름으로 자주 쓰이는 음절과 여자 이름으로 자주 쓰이는 음절이 있다. 통계적으로 '성','윤','찬'은 남자 이름의 끝에 자주 쓰이고, '미','옥','화'는 여자 이름의 끝에 자주 쓰인다. '석천', '종범', '준호'는 남자 이름으로 자주 쓰이고, '경미', '은영', '정화'는 여자 이름으로 자주 쓰인다.[11]

표 2. 한국 성씨 분포도
Table 2. Statistics for Korean Names

한국 성씨	
Frequency of Occurrence	Number
1,000,000+	5
100,000~999,999	46
10,000~99,999	53
1,000~9,999	53
100~999	73
10~99	26
1~9	30

다민족 국가인 미국은 성씨가 600만개가 넘는다고 한다. 그중 가장 흔한 성씨 7개가 차지하는 비율이 5% 정도이고 한사람만이 사용하는 성씨가 4백만 여명이라고 한다.

국내인의 성명은 성씨를 제외를 하면 1~2음절이 대부분이기에 실험에는 1~2음절에 해당하는 부분을 비교하여 남자, 여자를 판단하는 시스템을 구축하였다.

Ⅳ. 성명에 대한 성별 구분

국내인 성명의 성별 구분은 남녀 이름에 나타나는 특성을 이용한다. 한글 이름은 전화번호부[12]에서 추출하여 성별을 임의로 부여하였고 추출한 명단을 이용하여 수작업으로 남녀 한글 이름을 구분하고, 남녀 한글 이름의 특성 정보를 구축하였다. 남녀 한글 이름을 수작업으로 구분하였기 때문에 주관적인 견해가 포함될수 있다. 영어 이름은 NLTK Corpora[13]에서 제공하는 영어이름 Male, Female로 나누어져 있는 Names Corpus, Version 1.3을 사용하였다.

Naive Bayesian분류기[14]를 사용하여 3음절중 이름에 해당하는 1~2개의 음절을 입력 받아 실험데이터와 비교 후 남자가 자주 사용하는 이름과 여자가 자주 사용하는 이름을 구분하여 준다. 박준이나 정화같은 성씨와 이름이 2음절인 성명에 대해선 성씨까지 포함되어있는 2음절의 이름들은 전체를 비교 후 성별을 구분하려고 한다.

Ⅴ. 실험 및 분석

그림 3은 성명 인식 실험을 위해 수집한 데이터중 7000여개는 남자, 8000여개는 여자 이름으로 학습데이터는 남자, 여자를 합친 15000여개의 데이터를 무작위로 Shuffle하여 앞에서 5000개 뒤에서 5000개의 데이터로 학습시키고 실험을 하였다.


notaeho, notaehong, nohyeonho, nohyeong-gil, noh
bagseongjun, gimsincheol, baggeumseong, gimjung-
gimsiyeol, dountag, dojihyeong, dongho, bagjaeu,
lailhwan, lajongmun, yangdongmyeong, gangnamhyeo
ung, lyuhwiyeol, lyugeunhwang, lyusoyeol, lyuyeo
maenghyeonjae, chagibo, gimtaeug, myeongsangdeog
gimdaeheon, mundaehan, munbyeongju, munseongsu,
munbyeongcheon, munboin, munseonghwan, munseungj
munchangsig, muncheonsu, munhyoseog, bag-yeonglo
minbogi, minjihong, mintaegi, mingyeongbae, minm
bagchangsig, gimgiljung, baggang-yong, baggeon-u
baggiseong, baggiyun, baggijeong, bagnam-won, ba
bagbeomsu, bagbyeongmu, bagbyeong-il, bagbyeongj
bagseogbong, bagseongwan, bagseong-gyo, bagseong
bagseongjin, bagseongho, bagseonghun, bagsongju,
bag-wancheol, bag-yongsu, bag-yong-u, bag-yongji
bagjaeseog, bagjaeseong, bagjeongsu, bagjeong-og
bagjongcheon, bagjonghyeog, bagjonghwa, bagjongh
bagjin-yeong, bagjin-ug, bagjinhyeon, bagchandu,
bagtaehwa, baghajin, bagheongeol, bagheontae, ba
baghuiyeol, baghuijeong, baghuiho, baggeongyu, b
baggwangmyeong, baggwangbu, baggwangseo, baggwan
bagdaeu, bagdaejeong, bagdeogsun, bagdongsu, bag
bagmin-yeong, bagbyeong-geon, bagbyeongmin, bagb
gi, bagsangdon, bagsangmog, bagsangmun, bagsangs
bagsanghyeon, bagseogseon, bagseogjun, bagseogta
bagseong-yong, bagseong-won, bagseong-yun, bagse

그림 3. 실험 데이터
Fig. 3. Test Data

Naive Bayesian분류기로 성명 중 성씨를 제외한 부분의 영문자 1~8개를 입력받아 실험데이터와 비교를 하였다. 표 3, 표 4는 외국인과 국내인의 성별 결과이다.

표 3. 외국인 성별 분류 결과
Table 3. Classification for American Name's Gender

```
>>> classifier.classify(gender_features('nicolas'))
'mail'
>>> classifier.classify(gender_features('jessica'))
'female'
>>> classifier.classify(gender_features('hoult'))
'mail'
>>> classifier.classify(gender_features('amanda'))
'female'
```

표 4. 국내인 성별 분류 결과
Table 4. Classification for Korean Name's Gender

```
>>> classifier.classify(gender_features('inrak'))
'mail'
>>> classifier.classify(gender_features('hyjung'))
'female'
>>> classifier.classify(gender_features('nan'))
'female'
>>> classifier.classify(gender_features('hyeok'))
'mail'
```

표 5. 국내인 성명 분류 결과
Table 5. Statistics for Korean Gender Classification

성별	정확율
남	84%
여	88%
총합	86%

표 6. 미국인 성명 성별 분류 결과
Table 6. Statistics for American Gender Classification

성별	정확율
남	80%
여	84%
총합	83%

표 5는 국내인 성명의 경우 남자가 84%, 여자가 88%, 총합 86%, 표 6은 외국인 성명의 경우 남자가 80%, 여자84%, 총합 83%의 정확도를 확인 할 수 있었다. 하지만 사람 이름의 경우에는 남자와 여자가 동시에 같은 이름을 쓰는 경우가 있기 때문에 그러한 이름까지 전부 포함되어있고 일반 명사로 된 이름까지 포함된 실험을 하였다.

Ⅵ. 결론 및 향후 연구

본 논문은 국내인 성명에 대해 Python tool을 사용하여 Naive Bayesian 분류기를 사용하여 성명의 성별 분류를 실험하였다. 실험 결과로 국내인 성명은 남자 84%, 여자 88% 그리고 총합 정확율 86%로 분류가 되었고, 외국인 성명은 83%의 정확율을 보였다. 한국 사람이름에서도 성씨와 이름이 합쳐지면서 여자 이름이 되거나 남자 이름이 되는 경우가 있기 때문에 향후에는 사람 이름뿐만 아니라 성씨와 이름을 동시에 인식을 하여 사람 이름을 인식하여 성별 구분을 하는 연구도 필요하며, 요즘은 다문화 가정이 늘어남과 동시에 외국인들이 한국에 귀화하면서 생기는 한국 성씨와 외국 사람이름이 합쳐진 이름을 인식하여 성별을 구분하는 연구도 필요하다고 생각된다.

요즘은 사람 이름들이 일반 명사로 된 이름들도 많이 있기 때문에 일반 명사로 된 사람 이름과 성별을 조사하여 어떤 이름들이 남자로 혹은 여자로 사용되고 있는지, 성별 인식에 필요한 정보를 조사하고 연구해야 할 것이다.

References

[1] D. K. Lee, J. H. Kwon, “Social Search Algorithm considering Recent Interests of User”, Journal of Korean Institute of Information Technology, vol. 9, issue 4, pp. 187-194, Apr 2011.

[2] Y. H Kang, B. I. Kho, Y. H. Seo, "Unregistered Human Names Recognition and Sex Distinction", Dept. Computer Science, Chongbook Univ., 2004.

[3] K. H. Lee, J. H. Lee, M. S. Choi, K. C. Kim, "Korean Named Entity Recognition Based on Supervised Learning Using Named Entily Construction Princip", The 14th Annual

Conference on Human & Cognitive Language Technology, pp. 111-117, 2000.
[4] K. H. Lee, J. H. Lee, M. S. Choi, K. C. Kim,, "Study on Named Entity Recognition in Korean Text", The 12th Annual Conference on Human & Cognitive Language Technology, pp. 292-299, 2004.
[3] J. H. Lee, "The Role of Syntactic Cues in Pronoun Referential Resolution: The Effects of Number Cue and Gender Cue", Cognitive Science 15, pp. 25-33, 2004.
[5] Park, S.-B. and H.-G. Yoon. Determining the Gender of Korean Names for Pronoun Generation., World Academy of Science, Engineering and Technology 32, pp. 42-46. 2007.
[6] T. H. Kim, H. S. Lee, Y. S. Ha, M. H. Lee, S. H. Meang, "Proper Noun Extraction Using Data Sets",The 12th Annual Conference on Human & Cognitive Language Technology, pp. 11-18, 2000.
[7] H. M. Shin, "Gender Inference in Korean Newspaper Reading", The British & American Language & Literature Association of Korea 96, pp.161-177 , 2010.
[8] Humotion, GenderMotion, http://www.humotion.co.kr/, 2008.
[9] Erumy, "Name Analyze", 2008, http://www.erumy.com/nameAnalyze/eDefault.aspx.
[10] David L. Word, Charles D. Coleman, Robert Nunziata and Robert Kominski, "Demographic Aspects of Surnames from Census 2000", 2000
[11] Statistics Korea, Population Census 2003, http://kostat.go.kr, 2003.
[12] Korean Telephone Directory, "Telephone Directory", http://www.ktdc.co.kr, 1998.
[13] NLTK, "Natural Language Toolkit Development", https://code.google.com/p/nltk/, 2011.
[14] Stven Bird, Ewan Klein, and Edward Loper, "Natural Language Processing with Python", O'reilly, 2009.
[15] S. Moro, R. Laureano, and P. Cortez, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology", Proceedings of the European Simulation and Modelling Conference-ESM'2011, Guimaraes, Portugal, pp. 117-121, Oct. 2011.

저자 소개

임 명 재(종신회원)

• 을지대학교 의료IT마케팅학과 정교수

정 진 표(준회원)

• 을지대학교 의료IT마케팅학과 재학생

김 명 관(정회원)

• 1981년 3월~1985년 2월 : 숭실대학교 전자계산학과 학사
• 1985년 3월~1987년 2월 : 숭실대학원 전자계산학과 석사
• 1996년 9월~2004년 2월 : 숭실대학원 컴퓨터학과 박사
• 1989년 8월~1993년 2월 : 한국전자통신연구소 인공지능연구실 연구원

1993년 3월~2007년 2월 : 서울보건대학 컴퓨터정보과 부교수
2007년 3월~현재 : 을지대학교 의료IT마케팅학과 부교수
<주관심분야 : 인공지능, 자연어처리, 질의응답시스템, 시멘틱 웹>