미등록 이름 명사 인식 및 성별 구분

강유환° 고병일 서영훈
충북대학교 컴퓨터공학과
{eric°, kobi}@nlp.chungbuk.ac.kr yhseo@cbucc.chungbuk.ac.kr

Unregistered Human Names Recognition and Sex Distinction

Yuhwan Kang° Byeongil Ko Younghoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

본 논문은 사람 이름의 특성을 이용한 이름 인식과 이름의 성별 구분 방법에 대해 제안한다. 사람 이름을 묻는 질의문은 질의-응답 시스템에서 자주 나타난다. 모든 사람 이름을 사전에 등록하는 것은 어렵다. 경우에 따라서는 남녀 이름을 구분할 필요가 있다. 한국 사람 이름의 특성은 주로 3음절로 이루어져 있고, 성씨로 사용되는 음절의 수가 제한적이라는 것이다. 또한 이름에는 한자 독음이 많이 쓰이고, 남자 이름으로 자주 쓰이는 음절과 여자 이름으로 자주 쓰이는 음절이 있다. 이러한 특성을 이용하여 사람 이름 인식과 성별 구분을 수행한다. 일반 웹 문서에서의 실험 결과, 이름 인식의 정확률은 94%를 보였고, 남녀 이름 구분의 정확률은 98%를 보였다.

1. 서 론

질의-응답 시스템은 문서로부터 사용자가 원하는 해답을 찾아 제공해 주는 시스템이다. 질의 중에는 사람 이름을 묻는 질의, 회사나 기관과 같은 조직 이름을 묻는 질의, 지명을 묻는 질의, 시간이나 거리와 같은 단위를 묻는 질의 등이 있다. 문서로부터 해답을 추출하기 위해서는 사람 이름, 조직 이름, 지명, 단위 등과 같은 개체명을 인식하고 추출하는 방법이 필요하다.

개체명 인식에 대한 연구[1, 2, 3, 4, 5]는 정보검색에서 고유명사나 미등록어와 같은 색인어 추출을 위해 사용될 수 있으며, 질의-응답 시스템에서는 해답 추출을 위해 반드시 필요한 연구이다. 개체명 인식 연구는 크게 규칙 기반 연구[1, 2, 3, 4]와 통계 기반 연구[5]로 나눌 수 있다. 규칙 기반 연구는 개체명 인식을 위한 규칙을 수작업 또는 반자동으로 작성한 후 개체명을 인식하는 방법이다. 통계 기반 연구는 학습 코퍼스로부터 개체명 인식에 필요한 지식을 은닉 마르코프 모델, 최대 엔트로피 모델, 결정 트리 모델 등을 이용하여 학습한 후 개체명을 인식한다.

사용자 질의 중에는 사람 이름을 묻는 질의가 많기 때문에, 사람 이름 인식에 대한 연구는 질의-응답 시스템에서 사람 이름을 해답으로 추출하기 위해 필요하다. 그러나 기존 연구들에서는 사람 이름 인식에 대한 세부적 연구가 없었다. 질의 중에는 ' 영화 <너에게 나를 보낸다>의 여자 주인공은?' 과 같이 남녀 이름을 구분하여 묻는 경우가 종종 있다. 따라서 사람 이름을 남녀로 구분하여 인식할 필요가 있다.

본 논문에서는 한국 사람의 이름 특성을 이용하여 사람 이름을 인식하고 성별을 구분하여 인식하는 방법에 대해 제안한다. 사람 이름 인식과 성별 구분을 위해 성씨와 이름에 나타나는 음절의 특성 및 통계 정보를 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국 사람 이름의 특성에 대해 알아본다. 3장에서는 사람 이름의 특성을 이용한 이름 인식 및 성별 구분 방법에 대해 설명한다. 4장에서는 웹 문서를 대상으로 사람 이름 인식과 성별 구분 실험을 수행하고, 실험 결과를 분석한다. 5장에서는 결론 및 향후 연구에 대해 토론한다.

2. 사람 이름의 특성

한국 사람 이름의 95%이상이 3음절로 이루어져 있다. 사람 이름은 성씨와 이름으로 구분되며, 성씨로 사용되는

음절의 수가 제한적이다. 2000년 11월 현재 성씨의 종류는 286개이며, 100대 성씨가 차지하는 비율이 전체의 99.1%이다[6].

표 1 성씨의 비율(2000)

성씨	10대 성씨	20대 성씨	50대 성씨	100대 성씨
비율	64.4%	78.2%	94.4%	99.1%

이름에는 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.

3. 사람 이름 인식 및 성별 구분 방법

사람 이름은 성씨와 이름에 나타나는 음절의 통계 정보를 이용하여 인식한다. 통계 정보는 한국 통신에서 제공하는 전화번호부[7]로부터 외국인 이름을 제외한 사람 이름을 추출한 후 작성하였다.

성씨 사전은 상위 100개 미만의 성씨만을 이용하여 구축하였다. 모든 성씨를 이용하는 것은 비효율적이고, 잘못된 인식이 자주 발생하기 때문에 100개 미만의 성씨만을 사용하였다. 이름 사전은 사람 이름의 두 번째와 세 번째에 나타나는 음절의 종류와 각 빈도수를 계산하여 구축하였다. 또한 각 음절간의 바이그램 정보도 함께 구축하였다.

사람 이름의 성별 구분은 남녀 이름에 나타나는 특성을 이용한다. 전화번호부에서 추출한 이름은 대부분 남자 이름이고, 성별 구분이 표기되어 있지 않아 약 12,000여명의 이름을 충북대학교 학생 명단에서 별도로 추출하였다. 추출한 명단을 이용해 수작업으로 남녀 이름을 구분하고, 남녀 이름의 특성 정보를 구축하였다. 남녀 이름을 수작업으로 구분하였기 때문에 주관적 견해가 포함될 수 있다.

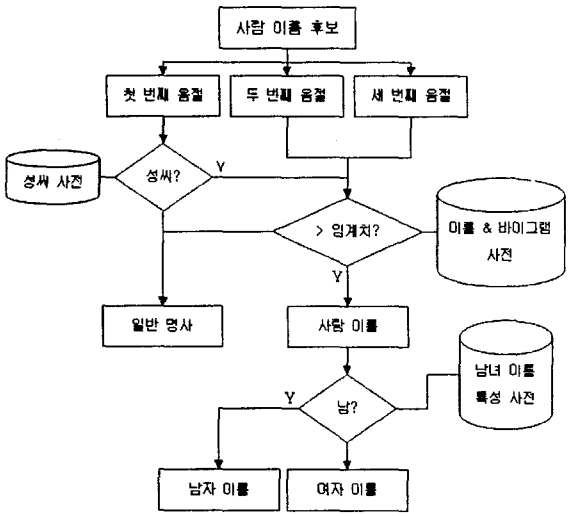

그림 1 사람 이름 인식 및 성별 구분

그림 1은 사람 이름 인식 및 성별 구분 순서도를 보여준다. 첫 번째 음절이 성씨이고, 두 번째 음절과 세 번째 음절의 통계 정보값이 임계치 이상이면 사람 이름으로 인식한다. 사람 이름으로 인식한 후에는 남녀 이름 특성 정보를 이용하여 남자 이름 혹은 여자 이름으로 구분한다.

4. 실험 및 분석

사람 이름 인식 실험을 위해 웹 문서로부터 2,190개의 3음절로 된 어절을 수집한 후 일반 명사와 사람 이름으로 구분하였다. 2,190개의 어절 중 사람 이름은 709개이다. 표 2는 사람 이름 인식의 정확률과 재현율을 보여준다.

표 2 사람 이름 명사 인식의 정확률과 재현율

구 분	정확률	재현율
성 능	94%	91%

사람 이름 인식의 정확률과 재현율은 다음과 같이 계산

한다.

- 정확률 = 시스템이 올바르게 인식한 사람 이름 수 / 시스템이 사람 이름으로 인식한 명사 수
- 재현율 = 시스템이 올바르게 인식한 사람 이름 수 / 총 사람 이름 수

사람 이름 중에는 사람 이름인지 일반 명사인지 구분하기가 모호한 경우가 있다. 예를 들면, '현미경' 은 사람 이름으로도 사용될 수 있지만 일반적으로는 명사로 사용된다. 객관적 실험을 위해 사람 이름인지 일반 명사인지 구분하기가 모호한 경우에는 형태소 분석용 사전에 수록된 명사인 경우에는 사람 이름으로 인식하지 않도록 하였다.

두 번째로 남녀 성별 구분 실험은 위 시험에서 사용한 709개의 사람 이름과 학습에 사용한 사람 이름 1,000개를 임의로 추출한 후 수행하였다. 남자 이름인지 여자 이름인지 객관적으로 구분하기 어려운 이름은 실험에서 제외하였다.

표 3 남녀 성별 구분 정확률

구분	정확률
성별	98%

남녀 성별 구분의 정확률은 98%로 다음과 같이 계산한다.

- 정확률 = 시스템이 올바르게 남자(여자)로 인식한 이름 수 / 시스템이 남자(여자) 이름으로 인식한 이름 수

5. 결론 및 향후 연구

본 논문에서는 사람 이름 인식 및 성별 구분을 위해 사람 이름의 특성을 이용하는 방법을 제안하였다. 실험 결과 사람 이름 인식의 정확률은 94%를 보였고, 남녀 성별 구분의 정확률은 98%를 보였다.

사람 이름을 묻는 질의문 중에는 외국인을 묻는 질의문도 많기 때문에 한국 사람의 이름뿐만 아니라 외국인의 이름을 인식하기 위한 연구도 필요하다.

현재 남녀 성별 구분은 성씨를 제외한 이름의 특성만을 이용하고 있다. 이름 중에는 이름만 봐서는 여자 이름이지만 성씨와 함께 사용할 경우에는 남자 이름인 경우가 있다. 마찬 가지로 그 반대인 경우도 존재한다. 이런 경우에는 성씨와 이름을 함께 고려할 필요가 있다. 따라서, 남녀 이름 구분을 위해 성씨와 이름 특성을 함께 고려하는 방법에 대한 연구가 필요하다.

참고 문헌

[1] 정래정, 김준태, " 고유명사의 출현 패턴을 이용한 색인의 성능 향상에 관한 연구" , 제8회 한글 및 한국어 정보처리 학술대회, pp. 68-72, 1996년

[2] 김태현, 이현숙, 하유선, 이만호, 맹성현, " 데이터 집합을 이용한 고유명사 추출" , 제12회 한글 및 한국어 정보처리 학술대회 , pp. 11-18, 2000년

[3] 이경순, 김재호, 최기선, " 한국어 질의응답시스템에서 개체인식에 기반한 대답 추출" , 제12회 한글 및 한국어 정보처리 학술대회, pp. 184-189, 2000년

[4] 이경희, 이주호, 최명석, 김길창, " 한국어 문서에서 개체명 인식에 관한 연구" , 제12회 한글 및 한국어 정보처리 학술대회, pp. 292-299, 2000년

[5] 황이규, 이현숙, 정의석, 윤보현, 박상규, " 개체명 구성 원리를 이용한 교사학습 기반의 한국어 개체명 인식" , 제14회 한글 및 한국어 정보처리 학술대회, pp. 111-117, 2002년

[6] 통계청, " 2000 인구주택총조사" , http://www.nso.go.kr, 2003년

[7] 한국통신, " 전국판 인명, 상호, 업종 CD번호부" , http://www.ktdc.co.kr, 1998년