韓國IT서비스學會誌
第15卷 第3號
2016年 9月, pp.115-125

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2016.15.3.115

한국어 소설에서 유정명사용 조사 기반의 인물 추출 기법*

박태근** · 김승훈***

A Character Identification Method using Postpositions for Animate Nouns in Korean Novels*

Taekeun Park** · Seung-Hoon Kim***

■ Abstract ■

Novels includes various character names, depending on the genre and the spatio-temporal background of the novels and the nationality of characters. Besides, characters and their names in a novel are created by the author's pen and imagination. As a result, any proper noun dictionary cannot include all kind of character names which have been created or will be created by authors. In addition, since Korean does not have capitalization feature, character names in Korean are harder to detect than those in English. Fortunately, however, Korean has postpositions, such as "*-ege*" and "*hante*", used by a sentient being or an animate object (noun). We call such postpositions as *animate postpositions* in this paper. In a previous study, the authors manually selected character names by referencing both Wikipedia and well-known people dictionaries after utilizing Korean morpheme analyzer, a proper noun dictionary, postpositions (e.g., "*-ga*", "*-eun*", "*-neun*", "*-eui*", and "*-ege*"), and titles (e.g., "*buin*"), in order to extract social networks from three novels translated into or written in Korean. But, the precision, recall, and F-measure rates of character identification are not presented in the study. In this paper, we evaluate the quantitative contribution of *animate postpositions* to character identification from novels, in terms of precision, recall, and F-measure. The results show that utilizing *animate postpositions* is a valuable and powerful tool in character identification without a proper noun dictionary from novels translated into or written in Korean.

Keyword : Information Extraction, Korean Novels, Character Identification, Postpositions for Animate Nouns, Korean Linguistic Feature

Submitted : June 20, 2016 1st Revision : July 20, 2016 Accepted : July 26, 2016

* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2016년도 문화기술 연구개발 지원사업으로 수행되었음.

** 단국대학교 응용컴퓨터공학과 교수, 교신저자

*** 단국대학교 응용컴퓨터공학과 교수

1. 서 론

정보 추출(Information Extraction)은 자연어 텍스트로부터 개체(Entity) 및 이벤트와 같은 중요한 정보들을 추출하는 작업이며(Küçük and Adnan, 2012), 개체명 인식(Named Entity Recognition)은 정보 추출의 일부분으로(Küçük and Adnan, 2012), 텍스트 내의 개체명을 발견한 뒤, 인명, 지명, 조직명과 같은 미리 정의된 클래스로 분류하는 작업이다(Nadeau and Kekine, 2007).

이러한 개체명 인식 기법들의 대부분은 규칙 기반 알고리즘 또는 기계학습 기반 기술을 활용하고 있는데(Nadeau and Kekine, 2007), 두 가지 기법의 단점은 줄이고 장점은 활용하고자 하는 하이브리드 기법들이 최근 제안되고 있다(Küçük and Adnan, 2012; Shaalan and Oudah, 2014).

그러나 개체명 인식 기법들에 대한 분석 연구(Nadeau and Kekine, 2007)에 따르면, 개발된 개체명 인식 기법들을 목표 텍스트 장르가 아닌 다른 텍스트 장르에 적용하는 것은 쉽지 않음에도 불구하고, 대부분의 연구들이 텍스트 장르와 도메인에 대한 영향을 크게 고려하지 않고 있다고 한다. 현재까지 제안된 대부분의 기법들은 신문 기사와 같은 텍스트로부터 개체명을 추출하는 것에 초점을 맞추고 있다.

본 논문은 한국어로 번역되거나 창작된 소설로부터 주요 등장인물, 인물 간 소셜 네트워크 및 시공간 배경 등을 컴퓨터가 자동으로 추출하는 것을 목표로 하는 연구의 일부분으로, 유정명사용 조사를 활용하여 한국어 소설로부터 인물명 및 등장인물(Character Names and Nominals)을 추출하는 기법을 제안하고자 한다.

소설에서 인물명 및 등장인물의 추출은 발화(인용기호 내의 문장)의 화자 식별(Speaker Identification)을 위해 필요하다(Elson and McKeown, 2010). 화자 식별이 되면, 인물 간의 소셜 네트워크를 파악하여 도서를 분류할 수도 있고(Elson et al., 2010), 화자의 성별, 나이 등을 파악하여 text-to-speech 기반 스토리텔링 시스템에서 화자에 어울리는 목소리로 책의 내용을 읽어 줄 수도 있다(Iosif and Mishra, 2014).

인물명은 “이사벨라”와 같은 고유 명사이며, 등장인물은 “아버지”와 같은 일반 명사이다. 신문 기사를 대상으로 하는 기존의 개체명 인식 기법들은 고유명사인 인물명만 추출하는데 반하여, 소설에서의 인물 추출에서는 고유명사인 인물명 뿐만 아니라 일반명사인 등장인물까지 추출하여야 한다(Elson and McKeown, 2010). 유정명사란 사람이나 동물 따위를 나타내는 명사를 의미하며, 유정명사 뒤에 붙을 수 있는 대표적인 조사로 ‘-에게’가 있다(Jeong, 2012). 본 논문에서는 유정명사 뒤에 붙을 수 있는 조사를 “유정조사”라고 간략히 표기하도록 한다.

본 논문에서 유정조사 기반의 인물 추출 기법을 제안하는 이유는 다음과 같다. 첫째, 작가들의 독창적인 문체로 작성되는 다양한 소설에 대하여 기계학습 기법을 적용할 수 있을 정도로 충분한 학습 데이터가 구축되기 이전에도 활용할 수 있는 인물 추출 기법이 필요하기 때문이다. 이것은 규칙 기반 개체명 인식 기법의 결과를 기계학습 기반 개체명 인식 기법의 입력으로 사용하는 하이브리드 기법의 등장 배경이기도 하다. 둘째, 소설의 배경 및 줄거리에 적합하도록 작가에 의해 창조되었거나 앞으로 창조될 모든 인물명을 포함하는 고유명사 사전의 구축이 어렵기 때문이다. 따라서 본 논문에서는 충분한 학습 데이터의 구축이 어려운 한국어 소설이라는 장르에 대하여 고유명사 사전에 의존적이지 않은 인물 추출 기법을 제안하고, 정확률(Precision), 재현율(Recall) 및 F-measure로 제안 기법의 성능을 분석하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구들을 소개하고, 제 3장에서는 한국어 소설에서 유정조사를 활용한 인물명 및 등장인물 추출기법에 대하여 기술한다. 제 4장에서는 성능 분석 결과를 살펴보고, 마지막으로 제 5장에서는 본 논문의 결론 및 향후 연구 방향을 정리한다.

2. 관련 연구

본 장에서는 소설에서의 인물명 및 등장인물 추출과 관련된 국내외 연구들을 소개한다.

Elson and McKeown(2010)에서는 영어 소설에서 발화자를 인식하기 위한 기법이 제안되었고, Elson et al.(2010)에서는 영어 소설에서 인물간 소셜 네트워크를 추출하는 기법이 제안되었다. 제안된 두 기법에서 공통적이면서 가장 먼저 수행되는 단계는 소설의 내러티브(Narrative) 부분에서 인물명(예 : "*Isabella*")과 등장인물(예 : "*her father*")을 인식하는 것이다. 신문 기사를 대상으로 하는 기존의 개체명 인식 기법들은, 이상의 두 기법과는 달리, 고유명사인 인물명만 추출한다는 차이점을 가진다. 이상의 두 기법에서는, 고유명사인 인물명을 추출하기 위하여 Stanford NER tagger를 사용하였고, 고유명사가 아닌 등장인물을 추출하기 위하여 정관사, 부정관사 및 소유격을 활용하였을 뿐만 아니라 상상속의 존재 등의 단어 목록을 제공하는 WordNet까지 활용하였다.

Iosif and Mishra(2014)에서는, 구텐베르크 프로젝트로부터 선택된 어린이 소설을 분석하는 다단계 시스템이 제안되었다. 이 연구의 목적은 Text-to-Speech(TTS) 기반 스토리텔링 시스템 개발을 위하여 소설 장르에 대한 분석 기술을 확보하는 것이다. 제안된 다단계 시스템에서는, 고유명사인 인물명을 추출하기 위하여 Stanford CoreNLP suite of tools를 활용하였고, 고유명사가 아닌 인간 또는 비인간 등장인물의 추출을 위하여 WordNet을 활용하였다.

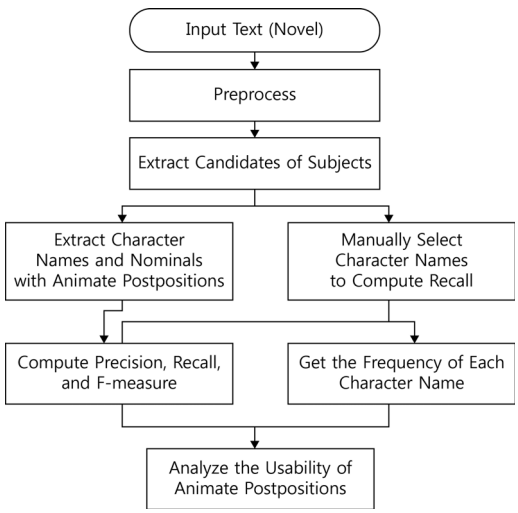
그러나 이상의 연구들은 대문자를 지원하는 영어로 작성된 소설을 대상으로 하고 있다. 대문자를 지원하지 않은 언어에서 고유명사를 식별하는 것이 영어에 비하여 월등히 어렵다는 사실은 이미 잘 알려져 있다. 하이브리드 기법을 제안한 연구(Küçük and Adnan, 2012)에 따르면, 어린이 소설을 대상으로 한 개체명 인식 실험에서, 대/소문자를 구분하는 텍스트에서의 F-measure는 81.61%였는데 반하여, 대/소문자 구분이 없는 텍스트에서의 F-measure는 74.68%에 불과하였다. 이러한 이유로, 아랍어를 위한 하이브리드 개체명 인식 시스템(Shaalan and Oudah, 2014)은 아랍어 단어를 영어로 번역한 다음, 번역된 단어가 대문자로 시작하면 이 단어를 고유명사로 표기하는 방법을 사용하기도 하였다.

아랍어나 한국어와 같이 대문자를 지원하지 않는 언어로 작성된 문서에서 고유명사인 인물명에 해당하는 개체를 추출하기 위하여, 많은 NER 시스템들은 실제 사람의 이름 또는 잘 알려진 사람의 이름 목록을 포함하는 고유명사 사전을 사용한다. 예를 들어, Seon et al.(2001)에서는 서울 전화번호부로부터 인물명을 수집하여 고유명사 사전을 만들기도 하였다. 그러나, 한 국가의 인물명으로만 구축된 고유명사 사전을 사용하는 NER 시스템은 다른 국가의 인물명 추출에 어려움을 겪을 수 있다(Küçük and Adnan, 2012).

한국어로 작성된 소설 또는 문헌 국역본에서의 개체명 추출 내용을 포함하는 연구는 많지 않다. Lee(2009)에서는, 19세기에 작성된 문헌의 국역본에 대하여 개체명 추출의 필요성을 역설하기는 하였으나, 연구 내용에서 개체명 추출은 수동으로 이루어졌다. Park et al.(2013)에서는, 한국어로 번역되거나 창작된 세 권의 소설에 대하여 등장인물간 소셜 네트워크의 구축을 위하여 인물명을 추출하기는 하였으나, 인물명 추출의 모든 단계를 자동화하지는 못했다. 구체적으로 서술하면, KAIST HanNanum 형태소 분석기와 조사 목록(예 : '-가', '-는', '-은', '-에게', '의' 등) 및 고유명사 사전을 활용하여 개체명을 자동 추출한 뒤, Wikipedia와 잘 알려진 인명사전에 대하여 추출된 인물명을 수동으로 교차 확인하는 작업을 수행하였다. 이상의 두 연구(Lee, 2009; Park et al., 2013)에서는 인물명을 추출하는데 있어서 사람의 개입을 필요로 하였기 때문에, 인물명 추출에 대한 정확률과 재현율 등의 성능 결과 값을 제시하고 있지 않다.

3. 한국어 소설에서 유정조사를 활용한 인물 추출 기법 및 성능 분석 방법

한국어로 번역되었거나 창작된 소설로부터 유정조사를 활용하여 인물명 및 등장인물을 추출하는 기법과 성능 분석 방법은 <Figure 1>과 같다.



〈Figure 1〉 Usability Analysis Procedure of Postpositions for Animate Nouns in Character Identification from a Novel

한국어 소설이 텍스트 파일 형태로 주어지면, 전처리 과정을 거친 뒤, 문장의 주어로 추정되는 모든 단어들을 자동으로 추출한다. 이러한 단어를 본 논문에서는 주어후보라 부른다. 다음으로, 소설 본문에서 유정조사와 함께 사용된 적이 있는 모든 주어후보를 인물명 및 등장인물(Character Names and Nominals)로 자동 추출한다. 이와는 별도로, 재현율 계산을 위하여 인물명에 해당하는 모든 주어후보를 수동으로 추출한다. 다음으로, 자동 추출된 인물명 및 등장인물 목록과 수동 추출된 인물명 목록을 이용하여, 정확률, 재현율, F-measure를 계산한다. 이에 추가로, 수동으로 추출된 인물명의 소설내 등장빈도를 계산한다. 마지막으로, 이상의 정보들을 활용하여, 한국어 소설에서 인물명 및 등장인물 추출에 대한 유정조사의 활용성 정도를 분석한다.

3.1 텍스트 전처리

소설은 내러티브(Narrative)과 발화(Utterance)로 구성된다. 발화는 소설 등장인물의 생각이 실제로 문장 단위로 실현된 것을 의미하며, 작가는 인용문 기호를 사용하여 특정 문장이 발화임을 표시한다. 내러티브는 소설의 줄거리를 이끌어 나가는 문장의 집합으로, 일련의 사건이 가지는 서사성을 1인칭 혹은 3인칭 관점에서 서술하는 문장들로 구성된다.

소설의 인물명은 내러티브에서 주어로 등장하지만, 많은 경우, 발화에서는 주어생략에 의해 인물명이 주어로 등장하지 않거나 대명사로 대체된다. 그러나 소설의 인물명은 발화에서 유정조사와 함께 등장할 수 있다. 이러한 이유로, 본 논문에서 주어후보를 추출할 때에는 내러티브에 해당되는 텍스트만 대상으로 하고, 유정조사를 이용하여 인물명과 등장인물을 추출할 때에는 내러티브 및 발화 전체 텍스트를 대상으로 한다. 따라서 텍스트 전처리 과정에서 내러티브에 해당되는 텍스트와 발화에 해당되는 텍스트로 원문 소설을 분리한다.

3.2 주어후보 자동 추출

한국어 문법에서 주격조사는 '-이/-가'이고 '-은/-는'은 보조사로 정의되어 있다. 하지만 소설을 포함하는 많은 문서에서, 주어가 될 수 있는 체언 뒤에 '-이/-가/-은/-는'을 붙여 주어로 사용하고 있다. 예를 들어, "해리가 말했다."와 "해리는 말했다."를 모두 사용하고 있다. 또한, 소설의 인물명들은 수차례부터 많게는 수백차례까지 내러티브에서 주어로 등장하기 때문에(Elson and McKeown, 2010; Elson et al., 2010), 받침 있는 인물명의 경우, ('-이', '-은') 조사 쌍 모두와 함께 내러티브 부분에 등장하거나, 받침 없는 인물명의 경우, ('-가', '-는') 조사 쌍 모두와 함께 내러티브 부분에

등장한다. 따라서 주어후보 자동 추출 단계에서는 내러티브 부분에서 (‘-이’, ‘-은’) 조사 쌍 또는 (‘-가’, ‘-는’) 조사 쌍 모두와 함께 사용된 적이 있는 모든 단어들을 주어후보로 추출한다.

이 단계에서 단어의 형태소를 분석하지는 않기 때문에, ‘많이’와 ‘많은’이라는 두 단어가 소설 내에 존재하는 경우, ‘많’이라는 한 글자가 주어후보로 추출될 수도 있다. 이와 같은 방법으로 주어후보를 추출하는 이유는, 작가에 의해 창조되는 다양한 인물명이 형태소 분석의 모호성에 의하여 주어에서 배제되는 경우를 막기 위함이다. 예를 들어, 꼬꼬마 형태소 분석기(Lee et al., 2010)로 “이사벨라는”을 형태소 분석하면, “[0/이사벨/일반명사], [3/라/(일반명사)+4/는/(주격조사)]”와 같은 형태소 분석결과가 얻어지며, “빌리는”을 형태소 분석하면, “[0/빌리/(동사)+2/는/(관형형 전성어미)]”와 같은 형태소 분석 결과가 얻어지는데, 본 논문에서는 “이사벨라”와 “빌리”와 같은 인물명을 모두 주어후보로 추출하기 위하여, 형태소 분석기를 사용하지 않는다.

3.3 유정조사를 이용한 인물명 및 등장인물 자동 추출

전 단계에서 추출된 주어후보에 대하여, 소설 본문에서 유정조사와 함께 사용되었는지 여부를 확인한다. 본 논문에서는, 어떤 주어후보가 유정조사와 함께 사용된 경우, 그 주어후보는 고유명사인 인물명이거나 고유명사가 아닌 등장인물로 추정한다. 이 때, 확인 대상이 되는 텍스트로는 소설의 내러티브뿐만 아니라 발화까지 모두 해당된다.

〈Table 1〉 List of Animate Postpositions

-hante(-한테)	-hanteseo(-한테서)
-hantero(-한테로)	-hanteneun(-한테는)
-ege(-에게)	-egeseo(-에게서)
-egero(-에게로)	-egeneun(-에게는)
-egen(-에겐)	-egekkaji(-에게까지)
-egedo(-에게도)	-egeseon(-에게선)
-egeseoneun(-에게서는)	

국어국립원 표준국어대사전에 포함된 366개의 조사 중에서 “(사람이나 동물 따위를 나타내는 체언 뒤에 붙어)”라고 표기되어 있는 유정조사와, 사전에는 등록되어 있지는 않지만, 소설 등에서 자주 사용되는 유정조사의 활용형태 목록은 <Table 1>과 같다.

그러나 <Table 1>에 나열된 유정조사와 함께 소설 본문에 등장하는 주어후보를 모두 추출한다면, 추출된 주어후보의 상당수가 대명사이거나 불특정한 사람을 나타내는 명사, 수사 또는 의존명사일 수 있다.

따라서 소설의 인물명 및 등장인물만 추출하고자 하는 본 단계에서는, 소설 본문에서 유정조사와 함께 사용되는 주어후보라고 하더라도,

1) 대명사(예 : ‘나’, ‘우리’, ‘그’, ‘그녀’ 등),
2) 불특정 명사(예 : ‘사람’, ‘남자’, ‘여자’ 등),
3) 집합명사(예 : ‘일가’, ‘가족’, ‘무리’ 등),
4) 복수형(예 : ‘사람들’, ‘남자들’, ‘여자들’ 등),
5) 수사(예 : ‘하나’, ‘둘’, ‘셋’ 등),
6) 의존명사(예 : ‘놈’, ‘명’, ‘분’, 등)에 해당하는 경우, 인물명 및 등장인물로 최종 선택되지 않도록 필터링한다.

3.4 인물명 수동 추출 및 등장빈도 계산

유정조사를 이용한 인물명 및 등장인물 추출 결과에 대한 재현율 계산을 위하여, 추출된 주어후보 중에서 인물명에 해당하는 것들을 수동으로 추출한다.

또한, 유정조사를 이용하여 자동 추출된 인물명과 그 인물명의 소설내 등장율과의 관계를 분석하기 위하여 주어후보에서 수동으로 추출된 인물명의 등장빈도를 계산한다. 소설내 인물명의 등장빈도를 계산하기 위하여, 국어국립원 표준국어대사전에 포함된 366개의 조사 중에서 인물명의 뒤에 붙을 수 있는 조사와, 발화에서 인물명이 불릴 때 사용되는 기호, 및 이들의 조합을 <Table 2>와 같이 선정하였다.

〈Table 2〉 List of Postpositions and Symbols for Getting the Frequency of Each Character Name

-ga(-가)	-gwa(-과)
-kke(-께)	-kkeseo(-께서)
-kkeopseo(-께옵서)	-neun(-는)
-da.(-다.)	-do(-도)
-rang(-랑)	-robuteo(-로부터)
-reul(-를)	-majeo(-마저)
-man(-만)	-mankeum(-만큼)
-bogo(-보고)	-buteo(-부터)
-siyeo(-시여)	-a(-아)
-a,(-아,)	-a.(-아.)
-a!(-아!)	-ya(-야)
-ya,(-야,)	-ya.(-야.)
-ya!(-야!)	-yamalro(-야말로)
-ege(-에게)	-egeda(-에게다)
-egero(-에게로)	-egeseo(-에게서)
-yeo(-여)	-yeo,(-여,)
-yeo.(-여.)	-yeo!(-여!)
-wa(-와)	-eurobuteo(-으로부터)
-eun(-은)	-eul(-을)
-ui(-의)	-i(-이)
-ida(-이다.)	-iraseo(-이라서)
-irang(-이랑)	-isiyeo(-이시여)
-iya.(-이야.)	-iyeo(-이여)
-cheoreom(-처럼)	-hante(-한테)
-hantero(-한테로)	-hanteseo(-한테서)
,(콤마)	.(마침표)
!(느낌표)	

소설 내 인물명의 등장율은 모든 인물명의 등장빈도 합에 대한 한 인물의 등장빈도 비율로 계산한다. 예를 들어, 인물명 A의 등장율이 1%라는 것은 소설의 전체 인물명의 등장빈도 합에 대하여 인물명 A의 등장빈도 비율이 1%라는 것을 의미한다.

3.5 정확률, 재현율, F-measure 계산

본 논문에서 정확률, 재현율 및 F-measure의 계산식은 다음과 같다.

$$\text{정확률} = \frac{\text{올바른 인물명 및 등장인물 수}}{\text{자동추출된 인물명 및 등장인물 수}}$$

$$\text{재현율} = \frac{\text{자동추출된 인물명 수}}{\text{수동추출된 인물명 수}}$$

$$F\text{-}measure = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}}$$

정확률은 유정조사를 이용하여 자동 추출한 인물명(고유명사) 및 등장인물(일반명사)에 대한 실제 인물명이나 등장인물의 비율로 계산된다. 이에 반하여, 재현율은 수동으로 추출된 인물명(고유명사) 중에서 얼마나 많은 인물명이 유정조사를 이용하여 자동 추출되었는지의 비율로 계산된다. F-measure는 정확률과 재현율의 조화평균으로 계산된다.

3.6 유정조사의 활용성 수준 분석

관련연구에서 언급한 바와 같이, 한국어로 번역되거나 창작된 소설에서 인물명을 추출하는 기존의 두 연구(Lee, 2009; Park et al., 2013)에서는 인물명의 추출에 사람의 개입을 필요로 하였기 때문에, 인물명 추출에 대한 정확률과 재현율 등의 성능 결과가 제시되어 있지 않다. 따라서 본 논문에서 제안하는 기법의 성능 수준을, 비록 한국어로 작성된 소설은 아니지만, 대/소문자를 구분하지 않는 텍스트로 표현된 어린이 소설을 대상으로 한 개체명 인식 연구(Küçük and Adnan, 2012)의 실험 결과와 비교한다.

다음으로, 본 논문에서 제안하는 기법의 활용성에 대하여 분석한다. 본 논문에서 제안하는 기법은 유정조사를 활용하기 때문에, 작가가 의인화를 사용하지 않는 경우, 자동 추출된 인물명과 등장인물은 모두 유정명사일 것으로 기대된다. 성능 분석 결과에서 상당히 높은 수준의 정확률이 얻어지는 경우, 본 논문에서 제안하는 기법이 어떻게 활용될 수 있는지에 대하여 분석한다.

4. 실험 및 결과

4.1 실험 대상 소설 목록

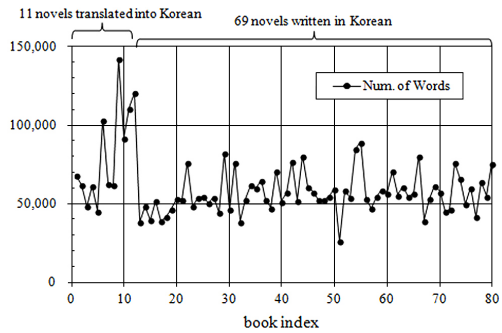
한국어 소설의 인물명 및 등장인물 추출에서 유정조사의 활용성 수준을 분석하기 위하여, 본 논문에서는 한국어 소설 80권으로 실험을 진행한다. 이 중에서, 11권의 소설은 한국어로 번역된 소설이며, 나머지 69권은 한국어로 창작된 소설이다. 실험에 사용되는 80권의 소설은 연구과제 수행을 위하여 확보된 소설들 중에서 임의로 선택하였다.

실험에 사용된 도서의 목록은 <Table 3>과 같은데, 한국어로 창작된 69권의 소설 이름을 모두 나열하기에는 공간이 부족하므로, <Table 3>에는 번역 소설 11권의 제목만 포함되어 있다.

〈Table 3〉 List of 80 Novels Translated into or Written in Korean for Experiment

	Book	Author
1	Blinder Instinct (사라진 소녀들)	Andreas Winkelmann
2	Nineteen Eight-Four (1984년)	George Orwell
3	The Detective is in the Bar (탐정은 바에 있다)	Azuma Naomi
4	A Little Princess (소공녀)	Frances Hodgson Burnett
5	Romance of the Three Kingdoms(part 2) (삼국지(중))	Lou Guanzhong
6	New Moon (뉴문)	Stephenie Meyer
7	O Zahir (오 자히르)	Paulo Coelho
8	Pride and Prejudice (오만과 편견)	Jane Austen
9	Breaking Dawn (브레이킹 던)	Stephenie Meyer
10	Twilight (트와일라잇)	Stephenie Meyer
11	Eclipse (이클립스)	Stephenie Meye
~80	69 novels written in Korean	

본 논문에서는, 약 50,000단어 정도로, 비슷한 단어 수를 가지는 한국어 소설들로 실험을 수행하려 하였으나, <Figure 2>에 보여지는 바와 같이, 100,000 단어 정도로 구성된 몇 개의 소설들도 실험 소설로 사용되었다. <Figure 2>의 x축은 소설의 인덱스를 나타내는데, 1번부터 11번까지의 소설이 한국어로 번역된 소설로서, 이 번호는 <Table 3>의 소설의 인덱스와 일치한다.



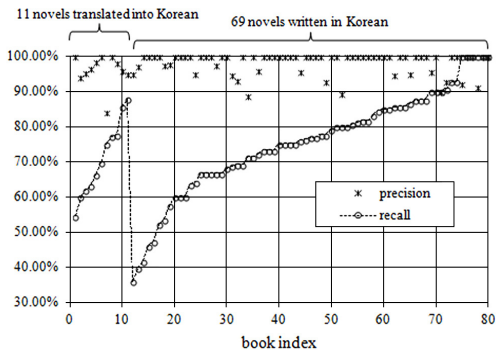
〈Figure 2〉 The Number of Words in Each Novel

4.2 실험 결과 및 활용성 분석

<Figure 3>은 80권의 한국어 소설에 대하여, 유정조사만을 사용하여 인물명 및 등장인물을 추출하였을 때의 정확률과 재현율을 보여준다. <Figure 2>와 동일하게, <Figure 3>에서 x축의 1번부터 11번까지는 번역 소설이며 12번부터 80번까지는 한국어 창작 소설이다.

<Figure 3>의 실험 결과에 따르면, 전체 80권에서 총 1,811개의 인물명 및 등장인물이 자동 추출되었으나 이 중에서 1,776개가 올바르게 추출된 것이어서, 전체 정확률은 98.07%로 계산되었다. 권당으로 바꾸어 표현하면, 권당 22.64개의 인물명 및 등장인물이 자동 추출되었고, 이 중에서 22.20개가 올바르게 추출된 인물명 및 등장인물이었다. 다르게 표현하면, 전체 80권의 책으로부터 인물명이나 등장인물이 될 수 없는 총 35개의 단어들이 추출되었는데, 이 단어들은 대부분 의인화되어 사용된 것들이었다. 예를 들면, “그 착한 목소리에게…”

또는 “지중해의 빛한테…”와 같은 문장에서 ‘목소리’와 ‘빛’이 의인화되어 유정조사와 함께 사용되었고, 그 결과 인물명 및 등장인물로 잘못 추출되는 결과가 초래되었다.



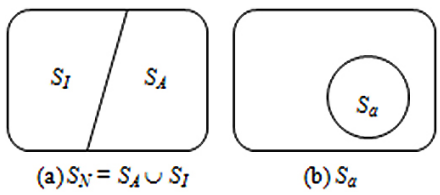
〈Figure 3〉 Precision and Recall in the Character Identification only with Animate Postpositions from 80 Korean Novels

<Figure 3>의 실험 결과로부터 재현율을 계산해 보면, 전체 80권으로부터 수동 추출된 총 1,433개의 인물명 중에서 1,007개의 인물명이 유정조사에 의해 추출되었으므로, 전체 재현율은 70.27%로 나타났다. 권당으로 표현하자면, 권당 17.91개의 인물명이 존재하지만, 유정조사로 찾아낼 수 있는 인물명은 12.59개였다.

이상의 정확률과 재현율로 F-measure를 계산해보면, F-measure는 81.88%가 된다. 비록 한국어로 작성된 소설은 아니지만, 대/소문자 구분이 없는 텍스트로 작성된 어린이 소설을 대상으로 한 하이브리드 개체명 인식 연구(Küçük and Adnan, 2012)의 F-measure 값이 74.68%인데 비하여 더 높은 성능을 보임을 확인할 수 있다.

그러나 <Figure 3>에서 각각의 소설별 재현율을 살펴보면, 한국어로 창작된 소설 중, 약 8권의 소설에서 재현율이 60%보다 낮게 나타났으며, 가장 낮은 두 개의 재현율은 36.00%와 39.66%에 불과하였다. 또한 재현율이 60% 이하인 8권의 소설 중에서 5권이 두 명의 작가의 소설로 나타났다. 이러한 사실로부터 유정조사를 활용하는 기법만으로는 작가의 문체에 따라 재현율의 편차가 커질 수 있음을 알 수 있다. 다르게 표현하자면, 작가가 유정조사를 즐겨 사용하지 않는 문체를 가지고 있는 경우, 제안하는 기법의 재현율은 낮아질 수 있다.

<Figure 4>는 소설 내의 명사의 집합 S_N과 유정명사의 집합 S_A, 무정명사의 집합 S_I 및 유정조사를 이용하여 추출한 인물명 및 등장인물의 집합 S_a의 관계를 보여준다.



〈Figure 4〉 Set of Nouns S_N, Set of Animate Nouns S_A, Set of Inanimate Nouns S_I, and Set of Extracted Character Names and Nominals S_a

소설 내의 모든 명사는 유정명사 또는 무정명사로 구분되며, 소설의 인물명(고유명사)과 등장인물(일반명사)은 유정명사의 집합 S_A에 속한다. 그러나 본 논문에서 유정조사를 활용하여 추출한 인물명과 등장인물의 집합은 <Figure 4>(b)의 S_a에 불과하다. 앞서 언급한 바와 같이, 작가의 문체에 따라 S_a의 크기가 결정되기 때문에, 어떤 작가가 쓴 소설이냐에 따라 제안하는 기법의 재현율 편차는 커질 수 있다.

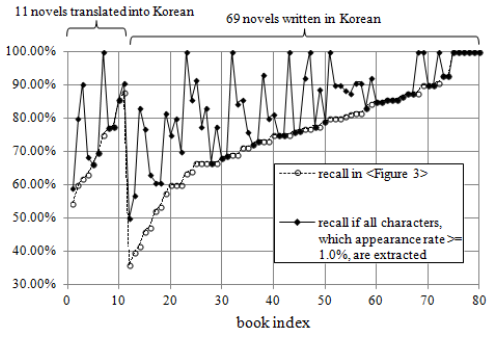
그러나 우리는 S_a의 정확률이 매우 높다는 것에 주목한다. S_a의 정확률이 100%에 가까운 경우, S_a에 속한 명사를 활용하여, 1) 이들과 관계를 가지거나 2) 이들이 사용된 문장과 유사한 패턴의 다른 문장에 등장하는, 명사를 인물명 또는 등장인물로 판단할 수 있을 것으로 생각한다.

예를 들어, 해리포터 시리즈에서 “해리”가 S_a에 속해있고 “헤르미온느”가 $S_N \setminus S_a$에 속해있으면서, 소설 본문에 “헤르미온느와 해리가”라는 문장의 일부가 존재하는 경우를 생각해 보자. 그러면, 우리는 “헤르미온느와 해리가”로부터 “해리”와 “헤르미온느”

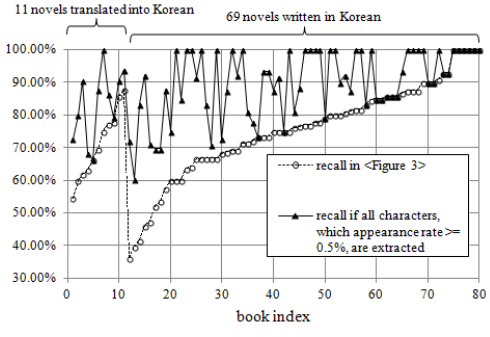
가 동등한 관계에 있다는 사실을 알 수 있고, 그 결과 "헤르미온느"도 인물명 또는 등장인물일 것이라 판단할 수 있다. 유사하게, S_a에 속해있는 명사의 소유격 패턴과 동사 패턴을 이용하여 $S_N \backslash S_a$에 속해있는 발견되지 못한 인물명 및 등장인물을 추출할 수 있을 것으로 생각한다.

이상의 아이디어를 구현하기에 앞서, S_a에 속한 인물명 및 등장인물을 활용하여, 각각의 소설별로 인물명의 등장율이 1% 이상 또는 0.5% 이상이면서 $S_N \backslash S_a$에 속해있는 모든 인물명을 추가로 발견하는 경우, 재현율이 얼마나 상승할 수 있는지 추정해 보면 다음과 같다. <Figure 5>와 <Figure 6>는 각각 등장율이 1% 및 0.5% 이상인 인물명을 모두 발견하는 경우의 80권의 재현율을 보여준다.

<Figure 5>로부터 재현율을 계산해 보면, 전체



〈Figure 5〉 Recall if All Characters, which Appearance Rate ≥ 1.0%, are Extracted



〈Figure 6〉 Recall if All Characters, which Appearance Rate ≥ 0.5%, are Extracted

80권으로부터 수동 추출된 총 1,433개의 인물명 중에서 1,119개의 인물명이 추출되어, 전체 재현율이 78.09%로 증가할 것으로 추정된다. 그리고 <Figure 6>로부터 재현율을 계산해 보면, 전체 80권으로부터 수동 추출된 총 1,433개의 인물명 중에서 1,200개의 인물명이 추출되어, 전체 재현율이 83.74%로 증가할 것으로 추정된다. 뿐만 아니라 <Figure 6>에서는, 각각의 소설별 재현율도 다섯 권을 제외하면 모두 70%보다 높게 나타났다.

<Figure 5>와 <Figure 6>의 결과로부터, 각각의 소설별 인물명의 등장율이 1% 이상 또는 0.5% 이상인 모든 인물명을 발견하는 경우의 F-measure를 계산해보면, 각각 86.94%와 90.34%가 된다.

이상의 성능 결과와 활용성 분석 결과로부터, 유정조사를 활용하여 한국어 소설로부터 인물명 및 등장인물을 추출하는 접근 방법은 매우 효율적인 방법일 뿐만 아니라, 새로운 인물 추출 기법 개발에도 유용하게 활용될 수 있음을 알 수 있다.

5. 결 론

본 논문에서는 충분한 학습 데이터의 구축이 어려운 한국어 소설이라는 장르에 대하여 고유명사 사전에 의존적이지 않은 인물 추출 기법을 개발하기 위하여, 인물명 및 등장인물 추출에 대한 유정조사의 활용성을 정확률, 재현율 및 F-measure로 분석하였다. 이를 위하여, 한국어로 번역된 소설 11권과 한국어로 창작된 소설 69권을 대상으로 실험을 수행하였다. 실험 결과, 단순히 유정조사를 사용하여 인물명과 등장인물을 추출하는 매우 간단한 방법만으로도 81.88%의 F-measure 값을 얻을 수 있었다. 또한 활용성 분석을 통해, 본 논문에서 제안하는 기법이 새로운 인물 추출 기법 개발에도 유용하게 활용될 수 있음을 알 수 있었다.

향후에는, 유정조사를 활용하여 추출된 인물명과 등장인물 정보를 기반으로, 제안하는 기법이 발견하지 못한 인물명 및 등장인물을 추가로 추출하는 기법에 대한 연구를 진행하고자 한다.

References

Elson, D.K. and K.R. McKeown, "Automatic Attiribution of Quoted Speech in Literary Narrative", *Procedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, 1013-1019.

Elson, D.K., N. Dames, and K.R. McKwown, "Extracting Social Networks from Literary Fiction", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, 138-147.

Iosif, E. and T. Mishra, "From Speaker Identification to Affective Analysis : A Multi-Step System for Analyzing Children' Stories", *the 3rd Workshop on Computational Linguistics for Literature*, 2014, 40-49.

Jeong, H., "A Cognitive Semantic Approach to Korean Particle Eygey", *Discourse and Cognition*, Vol.19, No.2, 2012, 133-152.

(정해권, "한국어 조사 '에게'의 인지의미론적 접근", *담화와인지*, 제19권, 제2호, 2012, 133-152.)

Küçük, D. and Y. Adnan, "A Hybrid Named Entity Recognizer for Turkish", *Expert Systems with Applications*, Vol.39, No.3, 2012, 2733-2742.

Lee, D.J., J.H. Yeon, I.B. Hwang, and S.G. Lee, "KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database", *Journal of KIISE : Computing Practices and Letters*, Vol.16, No.11, 2010, 1046-1050.

(이동주, 연종흠, 황인범, 이상구, "꼬꼬마 : 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구", *정보과학회논문지 : 컴퓨팅의 실제 및 레터,* 제16권, 제11호, 1046-1050.)

Lee, E.Y., "Named Entity Detection and Relation Extraction in the Personal Chronology of the 19th Century", *Journal of EONEIHAG*, Vol.53, 2009, 141-162.

(이은령, "19세기 문헌 국역본의 개체명 인식 및 관계 추출을 위한 기초 연구", *언어학*, Vol.53, 2009, 141-162.)

Nadeau, D. and S. Kekine, "A Survey of Named Entity Recognition and Classification", *Linguisticae Investigationes*, Vol.30, No.1, 2007, 3-26.

Park, G.M., S.H. Kim, and H.G. Cho, "Analysis of Social Network According to the Distance of Character Statements", *Journal of the Korea Contents Association*, Vol.13, No.4, 2013, 427-439.

(박경미, 김성환, 조환규, "소설 등장인물의 텍스트 거리를 이용한 사회 구성망 분석", *한국콘텐츠학회논문지*, 제13권, 제4호, 2013, 427-439.)

Seon, C.N., Y. Ko, J.S. Kim, and J. Seo, "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules", *In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium,* 2001, 229-236.

Shaalan, K. and M. Oudah, "A Hybrid Approach to Arabic Named Entity Recognition", *Journal of Information Science*, Vol.40, No.1, 2014, 67-87.

About the Authors

Taekeun Park (tkpark@dankook.ac.kr)

Taekeun Park received his B.S., M.S., and Ph.D. degrees in Computer Science and Engineering from POSTECH, Pohang, Korea in 1991, 1993, and 2004, respectively. He joined POSTECH PIRL in 1993 and moved to SK Telecom in 1996. From 2000 to 2001 and from 2001 to 2002, he worked for 3Com Korea and Ericsson Korea, respectively. In 2004, he joined in the department of Multimedia Engineering, Dankook University, Korea. He is currently on the faculty of the department of Applied Computer Engineering at Dankook University. His research interests include data processing, IoT, wireless/mobile communications, and distributed services.

Seung-Hoon Kim (edina@dankook.ac.kr)

Seung-Hoon Kim received his Ph.D. degree in Computer Science and Engineering from Pohang University of Science and Technology (POSTECH), Korea in 1998. Dr. Kim is currently a professor of Dept. of Applied Computer Engineering, Dankook University, Korea since 2001. From 1989 to 1990 he was a member of technical staff in Electronics and Telecommunications Research Institute(ETRI), Taejon, Korea. From 1991 to 1993 he was a member of technical staff in POSDATA, Seoul, Korea. His current research interests include data computing and networking, IoT, distributed systems, and etc.