한국정보전자통신기술학회논문지(jkiiect)'16-02, Vol.9 No.1
논문 15-09-01-075 http://dx.doi.org/10.17661/jkiiect.2016.9.1.075

한국어 소설에서 주요 인물명 인식 기법

김서희, 박태근, 김승훈*

A Recognition Method for Main Characters Name in Korean Novels

Seo-Hee Kim, Tae-Keun Park, Seung-Hoon Kim*

요 약 소설에서 주요 인물은 소설의 이야기를 전개하는 아주 중요한 역할을 담당하여 소설에서 없어서는 안 되는 중심인물을 의미한다. 기존의 인물명 인식 연구에서는 구축해놓은 인물명 사전을 통해 인물명을 인식하였고, 영어의 경우 대소문자 구별이 있으며 인물명과 함께 사용되는 단어를 활용하여 인물명을 인식하였다. 본 논문에서는 한국어 소설에서 용언, 규칙 및 가중치를 이용한 주요 인물명 인식 기법에 대해 제안한다. 먼저, 인물이 행할 수 있는 용언을 근거로 인물명 후보를 인식하고, 인식된 인물명 후보 중 인물명으로 사용될 수 없는 규칙에 해당되는 후보들을 제거한다. 문장에 나타나는 인물명 후보의 수에 따라 가중치를 부여하여 중요도를 계산하고, 중요도가 임계치 이상인 경우 주요 인물명으로 판단한다. 소설 300권을 대상으로 실험 결과 평균 85.97%의 정확도를 보였다. 인식된 주요 인물명은 향후 소설 내 등장인물 간 연관관계를 파악하거나 등장인물의 행위, 성향 등을 파악하는데 활용될 수 있다.

Abstract The main characters play leading roles in novels. In the previous studies, they recognize the main characters in a novel mainly based on dictionaries that built beforehand. In English, names begin with upper cases and are used with some words. In this paper, we propose a recognition method for main characters name in Korean novels by using predicates, rules and weights. We first recognize candidates for the characters name by predicates and propose some rules to exclude candidates that cannot be characters. We assign importances for candidates, considering weights that given by the number of candidates appeared in a sentence. Finally, if the importance of the character is more than a threshold, we decide that the character is one of main characters. The results from the experiments for 300 novels show that an average accuracy is 85.97%. The main characters name may be used to grasp relationships among characters, character's action and tendency.

Key Words : Data Mining, Korean Linguistic Feature, Korean Novels, Main Characters, Text Mining

1. 서론

소설에서 인물은 사건을 진행시켜 나가는 사람으로서 소설을 구성하는 가장 중요한 요소 중 하나이다. 인물은 역할, 중요한 정도, 특성과 같은 여러 분류 기준에 따라 나누어질 수 있다. 그중 중요한 정도에 따라 인물을 분류하면 주요 인물과 부수 인물로 나눌 수 있다. 주요 인물은 소설의 이야기를 전개하는데 중요한 몫을 담당하여 소설에서 없어서는 안 되는 중심인물을 의미하며, 동화 '신데렐라'의 '신데렐라', '새엄마' 등이 주요 인물에 해당된다.

기존의 인물명 인식 연구에서는 미리 구축한 인물명 사전을 기반으로 인물명을 인식하므로 사전에 등록되지 않은 인물명은 인식하기 어려우며, 사전에 많은 영향을 받는다. 또한 한국어의 경우 영어와 달리 대소문자의 구분이나 문장 작성 규칙이

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2015.
* Corresponding Author : Department of Applied Computer Engineering, Dankook University(edina@dankook.ac.kr)
Received February 5, 2016 Revised February 9, 2016 Accepted February14, 2016

없고, Mr, Jr 등 인물명을 나타내는 단어들이 없기 때문에 인물명 인식이 어렵다.

본 논문에서는 한국어 소설에서 용언, 규칙 및 가중치를 이용한 주요 인물명 인식 기법에 대해 제안한다. 먼저 인물명과 함께 사용 가능한 용언을 통해 문장을 추출하고, 인물명 제외 규칙을 적용하여 인물명 후보를 인식한다. 또한 문장 내에 등장하는 인물명 후보가 여러 개인 경우, 문장에 나타나는 행위의 주체를 판단하기 어렵기 때문에 문장에 등장하는 인물명 후보의 개수에 따라 가중치를 부여하여 소설에서의 주요 인물명을 인식한다.

2. 관련 연구

K. H. Lee 외 3명은[1] 개체명 사전, 결합 단어 사전을 이용한 규칙 기반의 한국어 개체명 인식 방법을 제안하였다. 개체명에는 인물명과 함께 지명, 날짜 등이 포함된다. 먼저 품사가 태깅된 신문 기사에서 어절 내 단어 및 주변 문맥을 확인한 뒤, 용언의 하위범주화 정보 및 개체명간 관계를 확인하였다. 해당 논문에서는 단계마다 개체명이라고 판단되는 수치를 부여하여 인물명 수치가 가장 높을 경우 인물명으로 인식하였다.

K. H. Bae 외 3명은[2] 계층형 개체명 사전을 구축하여 일정 및 개인 정보 관리에 관한 질의 문장을 대상으로 개체명을 인식하였다. 개체명 사전은 이름, 날짜, 장소 등으로 구성된 속성 개체명과 홍길동, 6일 등으로 구성된 인스턴스 개체명으로 이루어져 있다. 먼저 개체명 사전을 이용하여 입력된 문장의 단어가 어떤 개체명에 속하는지 분류하고, 판단이 애매한 단어의 경우 '이름 뒤에는 연락처가 온다.'와 같은 속성 개체명 패턴을 이용하여 인식하였다. S. K. Han은[3] 이러한 사전 기반의 인물명 인식 기법에 활용할 수 있는 인명사전을 구분하였다.

G. M. Park 외 2명은[4] 개체명 인식을 위해 어절의 복합 요소에 대한 형태소 분석을 수행하고, 그 결과 조사 '은, 는, 이, 가, 의, 에게'를 바탕으로 인명 후보를 선택하였다.

S. Morwal 외 2명은[5] Hidden Markov Model을 이용하여 인물명을 포함한 위치명, 조직명 등을 인식, 분류하는 방법을 제안하였고, E. Minkov 외 2명은[6] 영어로 작성된 이메일 텍스트를 대상으로 Mr, Jr 등 이름과 함께 사용되는 단어들의 특징을 이용하여 인물명을 인식하는 방법을 제안하였다. 또한 D. Maynard 외 4명은[7] 인물명 인식 방법 연구가 특정 도메인에만 집중되는 문제점을 해결하기 위해 영어로 된 다양한 텍스트 타입에서 인물명 인식 방법을 제안하였으며, R. Fu외 2명은[8] 중국어를 대상으로 영어-중국어 병렬 말뭉치를 이용한 인물명 인식 방법을 제안하였다.

그러나 한국어 인물명 인식에 관한 기존 연구들은 신문 기사와 정보 관리에 관한 문장을 대상으로 하고 있었으며 소설을 대상으로 한 연구는 찾을 수 없었다. 또한 대부분의 연구에서는 인물명 사전을 사용하는데, 사전을 구성하는 데이터의 양과 질에 따라 결과가 달라질 수 있다는 문제점이 있었다. 한국어의 경우 영어와 달리 대소문자나 문장 작성 규칙이 없고, Mr, Jr 와 같이 인물명을 나타내는 단어들이 없기 때문에 본 논문에서는 한국어 소설을 대상으로, 용언, 인물명 제외 규칙, 가중치를 이용한 주요 인물명 인식 기법에 대해 제안한다.

3. 주요 인물명 인식 기법

본 논문에서 제안하는 한국어 소설에서 주요 인물명 인식 기법은 그림 1과 같이 진행된다.

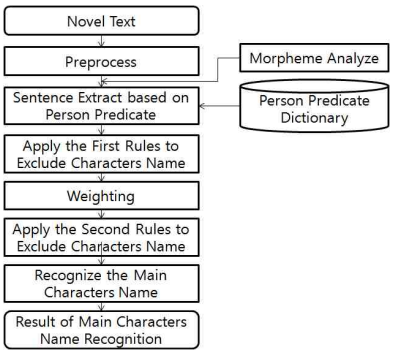

그림 1. 한국어 소설에서 주요 인물명 인식 기법
Fig. 1. A Recognition Method for Main Characters Name in Korean Novels

3.1 전처리

본 연구에서는 수집된 소설 본문에서 인물명을 인식하는데 불필요한 수식어 및 대화문을 제거하며, 분석에 용이하도록 문장을 분리하는 전처리 과정을 거친다. 소설 본문이 입력되면 괄호와 따옴표를 기준으로 그 안에 포함되어 있는 수식어 및 대화문을 제거한다. 또한 마침표, 물음표, 느낌표, 쉼표 등 문장을 구분하는데 사용되는 문장 부호들을 기준으로 문장을 분리한다.

3.2 인물 용언 사전 기반 문장 추출

용언은 동사와 형용사를 통틀어 이르는 말이다. 그 중 동사는 동작을 나타내는 품사로, 인물은 그 동작을 행하는 주체가 될 수 있으며 형용사를 이용하여 상태나 성질을 표현할 수 있다. 따라서 본 연구에서는 인물이 행하는 용언과 그 유의어로 구성되어 있는 사전을 구축한다. 기존 연구에서 사용한 사전의 경우, 인물명으로 구성되어 있으며 사전에 없는 인물명은 인식하지 못하기 때문에 분석 대상이 되는 문서에 따른 업데이트가 필요하다. 그러나 본 연구에서 구축한 인물 용언 사전은 한 번 구축되면 소설에 따른 영향을 거의 받지 않고 사용될 수 있다. 인물 용언 사전의 예는 표 1과 같다.

표 1. 인물 용언 사전의 예시
Table 1. Examples of Person Predicate Dictionary

약속하다	기뻐하다	이동하다	놀러가다
떠나다	판단하다	옮기다	이혼하다

소설 본문이 전처리 되면 본 논문에서는 꼬꼬마 형태소 분석기를 이용하여 모든 문장을 형태소 분석한다. 형태소 분석 된 문장에서 어떤 단어가 인물 용언 사전에 있는 경우, 해당 단어에 <iVV>태그를 부착하고 해당 문장을 추출한다. 예를 들어 '찰리가 넣어 둔 스파게티'라는 문장에서 '넣어'라는 단어가 용언 사전에 포함되어 있으므로, '찰리가 <iVV>넣어 둔 스파게티'와 같이 태그가 부착된다.

3.3 인물명 제외 규칙 1을 적용하여 인물명 후보 인식

한국어에서 인물명은 주어로 사용될 때 '은, 는, 이, 가'와 같은 조사를 부착한다는 특징이 있다. 따라서 본 연구에서는 문장에 '은, 는, 이, 가'로 끝나는 단어가 있으면 인물명 후보로 인식한다. 그러나 이렇게 인식 된 인물명 후보 중에는 인물이 아닌 단어가 다수 포함되어 있기 때문에 이를 제거해야 한다. 따라서 본 논문에서는 인물명 제외 규칙 1을 제안한다. 인물명 제외 규칙 1은 (a)부터 (h)와 같다. 규칙 적용 후, 제외되지 않은 인물명 후보에는 <S>태그를 부착한다. 예를 들어 '찰리가 <iVV>넣어 둔 스파게티'라는 문장에서 '찰리가'라는 단어가 '가'로 끝나며, 인물명 제외 규칙 1에 해당되는 것이 없으므로, '<S>찰리가 <iVV>넣어 둔 스파게티'와 같이 태그가 부착된다. 또 다른 예시로 '무엇이든 기꺼이 <iVV>먹어 <iVV>치우다'라는 문장에서 '기꺼이'는 '이'로 끝나지만, 인물명 제외 규칙 1의 (a)에 따라 인물명 후보에서 제외된다.

(a) 인물명 후보의 마지막 품사 태그가 '일반부사'인 경우
(b) 인물명 후보의 마지막 품사 태그가 '형용사+관형형 전성 어미'인 경우
(c) 인물명 후보의 마지막 품사 태그가 '보조 동사+관형형 전성 어미'인 경우
(d) 인물명 후보의 마지막 품사 태그가 '동사+관형형 전성 어미'인 경우
(e) 인물명 후보의 마지막 품사 태그가 '동사+의존적 연결 어미'인 경우
(f) 인물명 후보의 마지막 품사 태그가 '보조 동사+의존적 연결 어미'인 경우
(g) 인물명 후보 앞에 '~의'가 오는 경우
(h) 인물명 후보 앞에 '~을', '~를'이 오는 경우

3.4 문장 내 인물명 후보 개수 기반 가중치 부여

<iVV>태그와 <S>태그가 둘 다 포함된 문장에서 인물명 후보의 포함 개수에 따라 인물명 후보

에 가중치를 다르게 부여한다. 하나의 문장 내에 인물명 후보가 한 개인 경우, 해당 문장에서 동작을 행하는 주체는 하나이므로 인물명 후보 c_i의 가중치 w_i에 상수 $\widehat{w_1}$을 부여한다. 그러나 '신데렐라는 왕자님이 와서 기쁘다.'라는 문장에서 '신데렐라'와 '왕자님'이 인물명 후보인 경우, '오다'와 '기쁘다'라는 단어의 주체가 누구인지 판단하기 어렵다. 따라서 하나의 문장 내에 인물명 후보가 두개인 경우, 각 인물명 후보 c_i와 c_j의 가중치 w_i와 w_j에 동일한 가중치 $\widehat{w_2}$를 부여한다. 일반적으로 임의의 한 문장내 후보 개수가 n개이면, 인물명 후보 c_i에 대하여, $(i=1,\ldots n)$, 동일한 가중치 $\widehat{w_n}$을 부여한다. 즉, $w_1 = w_2 = \cdots = w_n = \widehat{w_n}$. 이제 부여되는 가중치 n-튜플 벡터 $\hat{\mathbf{w}}$는 식(1)과 같이 표현된다.

$$\hat{\mathbf{w}} = (\widehat{w_1}, \widehat{w_2}, \ldots, \widehat{w_n}), \text{ where } \widehat{w_1} \geq \widehat{w_2} \geq \cdots \geq \widehat{w_n} \quad (1)$$

3.5 인물명 제외 규칙 2를 적용하여 인물명 후보 인식

한국어에서 대명사는 사람의 이름을 대신 나타내는 품사로, 소설의 여러 인물들을 의미할 수 있기 때문에 해당 대명사가 어떤 인물을 지칭하는지 판단하기 어렵다. 가중치가 부여된 인물명 후보를 확인한 결과 '그녀', '그'와 같은 대명사가 높은 가중치를 가지고 있었다. 따라서 본 논문에서는 가중치가 부여된 인물명 후보의 형태소 분석 결과를 이용한 인물명 제외 규칙 2를 제안한다. 인물명 제외 규칙 2는 (i)부터 (p)와 같다. 예를 들어 '그는'의 경우 인물명 제외 규칙 2의 (n)에 해당하여 인물명 후보에서 제외된다.

(i) 일반 의존 명사+보조사
(j) 일반 의존 명사+'이다'
(k) 대명사+명사 파생 접미사+보조사
(l) 대명사+명사 파생 접미사+'이다'
(m) 대명사+주격 조사
(n) 대명사+보조사
(o) 대명사+'이다'
(p) 대명사

인물명 제외 규칙 2를 통해 수립된 각 인물명 후보의 중요도를 파악하기 위해 가중치를 사용한다. 즉, 인물명 후보 c_k의 중요도 $role(c_k)$는 소설 본문 전체의 각 문장에서 후보 c_k가 획득한 가중치의 합이 되며, 식(2)와 같이 계산한다.

$$role(c_k) = \sum_{\text{문장} s \in \text{소설본문}} \{w_i \mid c_k\text{가 문장} s\text{에서 후보} c_i\} \quad (2)$$

3.6 주요 인물명 인식

계산된 인물명 후보의 중요도를 기준으로 인물명 후보를 나열하고, 임계치 이상의 중요도를 갖는 인물명 후보를 주요 인물명으로 인식한다. 그림 2는 주요 인물명과 식(2)를 통해 계산된 중요도의 예시이다. 이 때, 임계치를 5라고 가정한다면 중요도 5를 갖는 '리'까지 총 24개가 주요 인물명으로 인식된다. 주요 인물명의 기준인 임계치는 중요도 값이나 주요 인물명의 수의 영향을 받으므로, 사용자의 요구에 따라 유동적으로 변할 수 있다.

```
에드워드 297        나/내 284          제이콥 200
앨리스 87           찰리 74            재스퍼 28
아빠 25             제인 15            로잘리 14
에밋 12             에스미 11          벤 11
칼라일 11           마이크 11          엄마 11
빅토리아 10         앤젤라 9           세스 9
샘 8                존 7               퀼 7
선생님 6            라일리 5           리 5
==================== PASS_THRESHOLD ====================
엠브리 4            브리 4             빌리 3
제시카 3            저레드 3           에밀리 2
돌 2                벨라 2             경찰 1
```

그림 2. 임계치에 따른 주요 인물명 인식 예시
Fig. 2. Example of Main Characters Name Recognition by Threshold

4. 실험 및 결과

4.1 실험 절차

본 논문에서는 다양한 장르의 한국어 소설과 외국 소설의 번역소설 등 총 300권을 대상으로 실험

을 진행하였다.

그림 3은 실험에 사용한 소설 중 '이클립스'를 대상으로 실험한 과정과 그 결과이다. 소설 '이클립스'의 본문이 입력되면 전처리 과정을 거친 뒤 형태소 분석을 한다. 인물 용언 사전, 인물명 제외 규칙 1, 2를 이용하여 인물명 후보를 인식하고, 가중치를 부여하여 중요도를 계산한다. 인식 된 인물명 후보 중 임계치 이상의 중요도를 갖는 인물명 후보인 '에드워드'부터 '리'를 주요 인물명으로 인식한다.

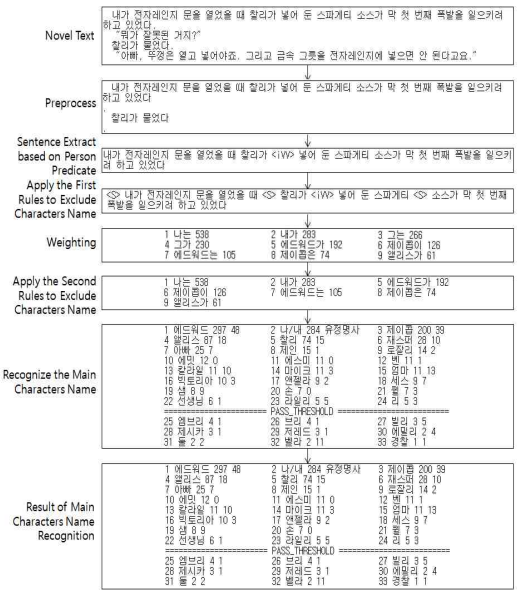

그림 3. 소설 '이클립스'에서 인물명 인식 과정 및 결과
Fig. 3. Main Characters Name Recognition Process and Result of Novel '이클립스'

4.2 실험 결과

본 논문에서 제안한 주요 인물명 인식 기법의 정확도를 확인하기 위해, 한국어 소설 300권을 대상으로 본 논문에서 권고하는 임계치를 적용하여 실험을 진행하고 그 결과를 분석하였다. 실험 결과는 표 2와 같다.

표 2. 인식된 주요 인물명의 수와 정확도

Table 2. Accuracy and Count of Recognized Main Characters Name

	Count	Accuracy
Maximum	51	100%
Minimum	6	55%
Mean	23.42	85.97%

주요 인물명 수는 최소 6명에서 최대 51명까지 인식되었다. 그 결과에서 주요 인물명 인식 정확도는 최소 55%에서 최대 100%까지 나타났으며, 300권 평균 85.97%를 보였다.

인식되는 주요 인물명의 수는 사용자의 요구에 따라 달라질 수 있다. 본 연구에서는 소설 300권을 대상으로 인식되는 주요 인물명의 수를 5개, 10개, 15개, 20개, 인식 가능한 최대 수로 나누어 그 정확도를 분석하였으며 인식되는 인물명의 수가 많아질수록 정확도가 떨어지는 것을 확인하였다.

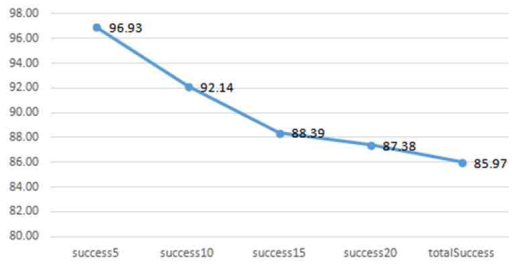

그림 4. 인식되는 주요 인물명의 수에 따른 정확도
Fig. 4. Accuracy Depending on Count of Recognized Main Characters Name

또한 본 연구에서는 소설 '이클립스'를 대상으로 인식한 주요 인물명과 '이클립스'의 위키피디아에서 제공한 주요 인물명을 비교하여 표 3을 통해 재현율을 확인하였다.

표 3. 소설 '이클립스'에서 인식된 주요 인물명의 재현율
Table 3. Recall Ratio of Recognized Main Characters Name in Novel '이클립스'

Main Characters Name from wikipedia	Recognized Main Characeters Name
에드워드	에드워드
벨라	나/내
제이콥	제이콥
칼라일	칼라일
아빠	아빠
재스퍼	재스퍼
로잘리	로잘리
빅토리아	빅토리아
제인	제인
앨리스	앨리스
라일리	라일리
빌리	(Under the Recall Ratio)
세스	세스
퀼	퀼
Recall Ratio (%)	92.86

5. 결론

본 논문에서는 한국어 소설에서 주요 인물명을 인식하기 위해 용언, 규칙 및 가중치를 이용한 기법을 제안하였다. 사람이 행할 수 있는 용언 및 유의어로 이루어진 인물 용언 사전을 구축하여 적용하고, 인물명이 될 수 없는 경우들로 이루어진 인물명 제외 규칙 1, 2를 적용하였다. 또한 문장 내 사용된 인물명 후보의 개수에 따라 가중치를 다르게 부여하고, 소설에서 인물명 후보의 중요도를 계산하였다. 마지막으로 인물명 후보의 중요도가 임계치 이상인 주요 인물명을 인식하였다. 본 논문에서는 제안한 기법의 정확도 확인을 위해 300권의 한국어 소설을 대상으로 실험을 진행한 결과, 평균 85.97%의 정확도를 보였다.

소설에서의 인물은 생각이나 행동, 인물과 인물 사이의 갈등을 통해 주제를 드러낼 수 있는 주체가 된다. 향후에는 인식된 인물명을 기반으로 등장인물의 연관관계, 행위 및 성향을 추출하는 연구를 진행할 계획이다. 이러한 정보들이 추출된다면 독자들이 원하는 소설을 선택하거나 소설의 내용을 이해하는데 도움을 줄 수 있을 것이다.

REFERENCES

[1] K. H. Lee, J. H. Lee, M. S. Choi, G. C. Kim, "Study on Named Entity Recognition in Korean Text", Proceedings of the 12th Annual Conference on Human and Cognitive Language Technology, pp. 292-299, October, 2000.

[2] K. M. Bae, S. H. Kim, Y. J. Ko, J. H. Kim, "An Efficient Named Entity and Topic Word Recognition Method Based on Named Entity Pattern in a Natural Language Interface", Journal of Korean Institute of Information Technology, Vol.12, No.1, pp. 121-129, Korean Institute Of Information Technology, January, 2014.

[3] S. K. Han, "A Comparative Study about Construction and the Service of the Domestic Biographical Database", Journal of Korean Library And Information Science Society (JKLISS), Vol.39, No.4, pp. 331-352, December, 2008.

[4] G. M. Park, S. H. Kim, H. G. Cho, "Analysis of Social Network According to The Distance of Characters Statements", Journal of The Korea Contents Association, Vol.13, No.4, pp. 427-439, April, 2013.

[5] S. Morwal, N. Jahan, D. Chopra, "Named Entity Recognition using Hidden Markov Model(HMM)", International Journal on Natural Language Computing (IJNLC), Vol.1, No,4, pp. 15-23. December, 2012.

[6] E. Minkov, R. C. Wang, W. W. Cohen, "Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text", Proceedings of the conference on Human Language Technology and Empirical Method in Natural Language Processing, pp. 443-450, Association for

Computational Linguistics, 2005.
[7] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, Y. Wilks, "Named Entity Recognition from Diverse Text Types", In Recent Advances in Natural Language Processing 2001 Conference, 2001.
[8] R. Fu, B. Qin, T. Liu, "Generating Chinese named entity data from parallel corpora", IJCNLP, 2011.

저자약력

김 서 희(Seo-Hee Kim) [학생회원]

- 2015년 2월 : 단국대학교 멀티미디어공학과 (학사)
- 2015년 3월 ~ 현재 : 단국대학교 대학원 컴퓨터학과 (석사과정)

<관심분야> 데이터컴퓨팅, IoT

박 태 근(Tae-Keun Park) [정회원]

- 1991년 2월 : POSTECH 컴퓨터공학과 (학사)
- 1993년 2월 : POSTECH 컴퓨터공학과 컴퓨터 통신 및 네트워크 전공 (석사)
- 2004년 2월 : POSTECH 컴퓨터공학과 컴퓨터 통신 및 네트워크 전공 (박사)
- 2004년 9월 ~ 현재 : 단국대학교 멀티미디어공학과 교수

<관심분야> 정보통신, 정보시스템

김 승 훈(Seung-Hoon Kim) [정회원]

- 1985년 2월 : 인하대학교 전자계산학과 (학사)
- 1989년 8월 : 인하대학교 대학원 전자계산학과 (석사)
- 1998년 2월 : 포항공과대학교 대학원 컴퓨터공학과 (박사)
- 2001년 9월 ~ 현재 : 단국대학교 응용컴퓨터공학과 교수

<관심분야> 정보통신, 데이터컴퓨팅