

효율적인 한국어 파싱을 위한 최장일치 기반의 형태소 분석기 기능 확장

이현영*, 이종석*, 강병도**, 양승원*

요약

한국어는 문장 구성소의 생략과 수식 범위가 자유롭기 때문에 파싱보다는 형태소 분석 단계에서 처리하면 좋은 경우가 있다. 본 논문에서는 파싱의 부담을 덜어 줄 수 있는 형태소 분석기의 기능 확장 방안을 제안한다. 이 방법은 미지어의 추정, 복합 명사 및 복합동사의 처리, 숫자 및 심볼의 처리에 의해 여러 형태소 열이 하나의 구문 범주를 가질 때 이것을 최장일치 방법으로 결합하고 의미 자질을 부여하여 하나의 구문 단위로 처리하는 것이다. 제안한 형태소 분석 방법은 불필요한 형태론적 모호성이 제거되고 형태소 분석 결과가 줄어들어 태거 및 파서의 정확률이 향상되었다. 또한, 실험을 통해 파싱 트리는 평균 73.4%, 파싱 시간은 평균 52.9%로 줄었음을 보인다.

키워드 : 문장 구성소, 형태소 분석기, 최장일치법, 파서

Functional Expansion of Morphological Analyzer Based on Longest Phrase Matching For Efficient Korean Parsing

Hyeon-yoeng Lee*, Jong-seok Lee*, Byeong-do Kang**, Seung-weon Yang *

Abstract

Korean is free of omission of sentence elements and modifying scope, so managing it on morphological analyzer is better than parser. In this paper, we propose functional expansion methods of the morphological analyzer to ease the burden of parsing. This method is a longest phrase matching method. When the series of several morpheme have one syntax category by processing of Unknown-words, Compound verbs, Compound nouns, Numbers and Symbols, our method combines them into a syntactic unit. And then, it is to treat by giving them a semantic features as syntax unit. The proposed morphological analysis method removes unnecessary morphological ambiguities and decreases results of morphological analysis, so improves accuracy of tagger and parser. By empirical results, we found that our method decreases 73.4% of Parsing tree and 52.4% of parsing time on average.

Keywords : sentence element, morphological analyzer, longest phrase matching method, parser

1. 서론

※Corresponding Author : Seung-weon Yang

Received : March 21, 2016

Revised : June 01, 2016

Accepted : June 30, 2016

* Woosuk University Game & Contents

Tel: +82-10-4626-2981, Fax: +82-63-290-1459

email: yangy123@woosuk.ac.kr

** Daegu University Computer & Information Technology

■ This Research was financially supported by Daegu University 2013.

형태소 분석은 어절을 구성하는 각각의 형태소들을 인식하는 단계이다[1]. 그런데 한국어는 특성상 생략과 도치가 빈번하고 각종 기능어 및 접사가 많이 쓰이며 용언의 활용이 다양하다. 따라서 형태소 분석을 수행하면 불필요한 후보들이 많이 생성되어 형태소 분석 결과를 이용하는 태거나 파서 등 여러 응용 시스템에는 부담으로 작용한다. 그래서 최근에는 응용 프로그램의 성

격에 맞도록 형태소 분석기의 기능을 확장하거나[2,3], 형태소 분석기를 사용하지 않고 태깅을 하는 시스템[7]이나 형태소 분석 시스템을 경량화하려는 연구[8]가 이루어지고 있다.

파싱의 관점에서 형태소 분석기의 기능을 확장한 대표적인 연구로는 [2]와 [3]이 있다. [2]에서는 “본용언”과 “보조용언”의 결합 관계를 연구하여 하나의 단위로 묶일 수 있는 것들을 묶고 의미자질을 추가하여 하나의 본용언으로 처리하였다. [3]은 [2]의 연구를 토대로 하나의 구문적 단위로 묶일 수 있는 보조용언과 의존명사나 의사조사를 포함한 형태소 열을 “구문 형태소”라고 정의하고 이를 체계화하였다. 이 방법은 평균 형태소 분석 결과가 41.30개에서 20.38개로 줄어들었고 파싱의 성능도 향상되었다. 그러나 이들 연구에서는 구문 모호성의 많은 부분을 차지하는 복합 명사에 대한 연구가 없었으며 명사구를 이룰 수 있는 단어열 수사나 미등록어, 심볼 등에 대한 고려는 이루어지지 않았기 때문에 여전히 파싱에게는 부담으로 작용한다.

따라서 본 연구는 [3]에서 다루지 않았던 복합 명사와 단어열 수사, 미지어, 심볼 등에 대한 연구를 진행하고 이들 중에서 여러 형태소가 모여 하나의 구문적 의미를 가지는 형태소들을 최장일치 방법으로 묶어 하나의 구문 기능을 가지도록 형태소 분석 단계에서 처리해 주면 파싱의 부담이 줄어들 뿐 아니라 파싱 결과도 정확해질 수 있다. 이와 같이 본 논문에서는 파싱의 부담을 줄일 수 있도록 결과를 생성해 주는 형태소 분석기의 기능 확장을 제안한다.

2. 최장일치법

최장일치법은 단어를 이루는 부분 문자열 집합을 구할 때 그 단어를 이루고 있는 형태소들의 집합 중에서 가장 긴 형태소를 우선적으로 선택하여 처리 하는 것이다. 예를 들어 “사회”, “복지”, “위원회”, “사회복지”가 표제어로 등록되어 있는 상태에서 “사회복지위원회”를 최장일치 방법으로 형태소 분석하면 “사회복지/명사+위원회/명사”로 분석된다.

그러나 본 논문에서는 사전을 검색할 때 뿐만 아니라 파싱의 관점에서 형태소 분석된 결과를

이용할 때도 최장일치법을 사용한다. 즉, 형태소 분석된 결과들 중에서 여러 형태소 열이 모여 하나의 구문 기능을 가질 경우, 이를 최대한으로 묶고 하나의 구문 기능과 속성 값을 할당한다. 앞의 예에서처럼 “사회복지위원회”가 “사회복지/명사+위원회/명사”로 분석이 되면 파싱에서는 “사회복지”와 “위원회”의 결합관계를 검사하기 위해서 파싱 문법을 기술하고 복잡한 처리 과정을 거쳐야 한다. 또한, 단어 앞과 뒤에 오는 정보에 의해 많은 구문 모호성이 발생하게 된다. 그러나 형태소 분석 후처리에서 “사회복지”와 “위원회”를 결합하여 “사회복지위원회/명사”로 분석해 주면 파싱 단계에서의 문제점이 해결될 뿐만 아니라 파싱 결과도 더 좋아질 수 있다. 이런 이유로 파싱의 관점에서 형태소 분석기를 재해석하고 파싱 단계에서 처리하는 것보다 형태소 분석 단계에서 처리하면 좋은 것들을 형태소 분석 후처리에서 처리하는 방안을 제안한다.

3. 형태소 분석기의 기능 확장

3.1 미등록어 처리

형태소 분석에서는 사전에 등록되지 않은 인명이나 복합 명사 등 평균 3.5%를 차지하는 미등록어를 분석할 때 많은 문제가 발생한다[4]. 이를 해결하기 위한 방안으로 미등록어를 사전에 수록하는 추세이지만 많은 단점이 있다. 따라서 본 연구에서는 출현 빈도가 높은 단어들은 사전에 수록하는 것을 원칙으로 하고 자주 사용되지 않는 단어들만 미등록어 처리를 수행한다.

미등록어는 대부분 접두사, 접미사, 조사, 어미 등과 같은 기능어가 동반되는 경우가 많다. 따라서 형태소 분석에 실패한 미등록어가 발견되면 미등록어 처리 모듈을 통해 미등록어의 앞뒤에 나타나는 기능어 정보를 이용하여 미등록어를 분석한다.

1) 체언형 미등록어

규칙:[접두사+] 단어 [+접미사] [+조사] [+어미]

예) 노무현씨를:노무현/명사+씨/접미사+를/조사
햇과일:햇/접두사+과일/명사

앓은뱅이:앓은/?+뱅이/접미사 =>사전에 등록

‘앉은뱅이’에서는 ‘뱅이’라는 접미사 정보를 이용하여 ‘앉은’을 명사로 유추할 수 있는데 ‘앉은’은 명사가 아니다. 따라서 ‘앉은뱅이’는 사전에 등록하는 것을 원칙으로 한다.

2) 용언형 미등록어

규칙 : [접두사+] 단어 [+접미사] [+선어말어미] [+조사 또는 어미]

예) 설익은, 노무현답다

‘갓구운’에서는 ‘갓’이라는 접두사 정보를 떼고 ‘구운’을 형태소 분석하면 용언으로 분석된다. ‘노무현답다’에서는 용언화 접미사 ‘답’이 붙어 ‘노무현답’이 용언의 어간으로 처리된다. 이와 같이 미등록어는 기능어들의 결합관계를 테이블로 작성하여 분석을 할 수 있다. 또한 조사나 어미에도 최장 일치의 방법을 적용하여 조사는 1,277개, 어미는 1,145개로 확장하여 사용한다.

그러나 미등록어가 나타난 앞뒤의 단어 정보나 우선순위 규칙을 이용하여만 미등록어를 추정할 수 있는 경우도 있다. 예를 들면

문장 : “복수어 단위나 신조어와 같은 미등록어를 추정하기가 쉬워진다.”

‘복수어 단위나 ...’는 명사 ‘단위’ 앞에 ‘복수어’가 있다. 따라서 ‘복수어’는 명사를 꾸며주는 관형사이거나 명사 ‘단위’와 결합하는 복합 명사이어야 한다. 그런데, “복수어”는 관형형 어미를 갖고 있지 않으므로 관형사가 아닌 명사로 추정한다. ‘신조어와’는 ‘신조어+와’와 ‘신조어+와’로 분석되지만 ‘같다’의 단위 정보로 ‘-와 같다’가 있으므로 ‘신조어’를 명사로 추정할 수 있다. 또한, ‘...같은 미등록어를 추정하기가...’에서는 의사조사 ‘와 같은’과 관형형 어미 정보 및 목적격조사 ‘를’을 이용하여 ‘신조어와’와 ‘미등록어를’ 바로 명사로 추정한다.

복합 동사는 용언의 특성을 갖는 미등록어이다. 복합 동사를 분류하면 ‘해놓다’와 같이 동사와 보조 용언이 통합되어 한 단어처럼 사용되는 것과 ‘밥먹다’와 같이 조사의 생략으로 인하여 체언과 용언이 통합된 것으로 분류된다. 이러한 것들은 특별한 규칙이 없기 때문에 사전에 등록하여 처리하는 것으로 한다.

3.2 숫자의 처리

숫자의 유형에서 정보가 될 수 있는 특징들을 추출하여 분석을 하면 보다 정확한 숫자 처리가 가능하다. 예를 들면 분수의 ‘/’, 자릿수의 ‘,’, 소수점 ‘.’, 연도의 ‘’, ‘BC’, ‘AD’, 전화번호의 ‘()’, ‘-’ 등의 정보를 이용한다. 또한, 숫자와 더불어 수학 기호가 연속하여 결합되어 있을 때는 하나의 수식으로 처리하고 숫자와 날짜 및 시간을 나타내는 단위가 연속되어 나타날 때에도 이 형태소들을 결합하여 하나의 단위로 처리한다.

또한, 수사는 ‘만’ 단위씩 띄어 쓰는 특성에 의해 한 어절을 넘기는 경우가 많다. 수사 분석은 수사 어절이라고 분석이 된 첫 어절부터 단위 명사 어절, 일단위 수사 어절, 서수사, 조사와 결합된 어절까지를 분석의 한 단위로 한다. 수사는 첫 문자가 ‘하, 한, 둘, 일, 이,’ 등으로 한정되어 있으므로 첫 문자를 이용하여 수사일 가능성을 판단한다. 예를 들어 “이억 이백삼십만 오천만은”이라는 문장의 분석과정은 다음과 같다.

문장 : 이억 이백삼십만 오천만은

1. 자연적 순위 결정 : 이억 이백삼십만 오천 (수사)

형태소 추출 : 만(수사, 조사)

2. ‘만’의 가능한 어절 순위는 단위인 ‘만’의 어절 순위보다 작다. 따라서, ‘만(조사)’가 형태소로 선택된다.

3. 분석 결과 : 이억 이백삼십만 오천(수사) + 만(조사) +은(조사)

3.3 심볼의 처리

문장 부호가 포함된 단어에서 영어 문자와 숫자, 문장 부호는 미리 분리해서 처리한 후에 한글 부분을 추출한다. 형태소 분석시에 기본적으로 처리해야 할 영문자, 숫자, 문장 부호를 포함한 단어의 유형은 다음과 같다.

- 영문자 + 조사/어미
예) Win10으로 UP하다
- 부호 + 단어 + 부호
예) ‘GUI’는, ‘UX’는
- 한글 + 부호 + 단어 + 부호
예) 서울(Seoul)은, 학생(學生)은

한글 이외의 문자들은 전처리 단계에서 한글과 분리하여 독립적인 단어로 간주하여 분석하는데, 이 때 문제가 되는 것은 뒤에 조사나 ‘-하다’류의 한글이 붙는 단어와 체언과 조사 사이에 괄호가 삽입되는 단어이다. 이러한 단어들은 한글 부분을 분석한 후에 다시 결합해 주어야 한다. 전처리 단계를 거쳐 추출된 한글 단어는 조사가 분리되면 체언 분석 과정에서 체언인지 용언과 어미 뒤에 조사가 결합한 경우인지를 구분하여 분석하고, 어미가 분리될 경우에는 용언 분석과정에서 용언인지 서술격 조사 ‘이’가 어미와 결합한 체언인지를 구분하여 분석한다.

3.4. 복합 명사, 복합동사의 처리

복합 명사와 명사 연결형의 명사구를 명사 연결 구성이라 하여 함께 다룬다. 또한, 언어를 보다 효율적으로 처리하기 위해 형태소 분석 과정에서 복합 어휘를 단일 명사 단위로 분해하는 것이 아니라 하나의 의미 단위로 처리하여 구문 분석의 부담을 줄이도록 한다. 본 논문에서 사용하는 명사들의 연결 구성은 다음과 같다.

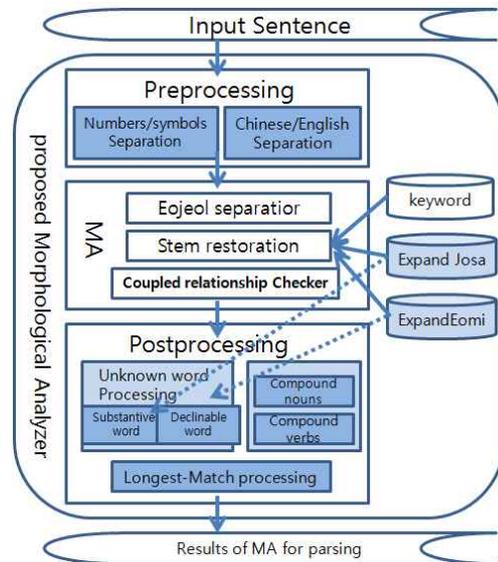
- (1) 명사가 나열된 형태
 - 예) 정보 검색 방법 => 정보검색방법
- (2) 조사가 생략됨으로써 발생하는 명사구
 - 예) 칠수(의) 얼굴 => 칠수얼굴
- (3) 접속조사 “와/과/또는” 등의 문장소에 의해 접속된 명사구
 - 예) ((LG 냉장고)와 (삼성 냉장고))
 - (칠수와 (영희의 부모)) <=> ((칠수와 영희의) 부모)
- (4) 소유격조사 “의”에 의해 접속된 명사구
 - 예) 칠수의 얼굴 => 칠수의얼굴
- (5) 관형사: 가장 가까운 명사를 수식
 - (그 지진의 강도) - X
 - ((그 지진)의 강도) - O
- (6) 명사 + 관형격 조사: 뒤에서부터 수식
 - ((궁전의 공주)의 방)) - X
 - (궁전의 (공주의 방)) - O
- (7) 의존명사 + 관형격 조사: 의존 명사는 관형어의 수식을 먼저 받음
 - (색다른 (곳의 느낌)) - X
 - ((색다른 곳)의 느낌) - O

복합 동사는 [3]에서 제안한 것들을 그대로 수용하며 “명사+용언”의 형태를 추가한다. “명사+용언”의 형태는 “잠자다”, “밥먹다”와 같은 구조는 띄어쓰기 오류를 범한 것이지만 실생활에서도 자주 사용되는 경우에는 사전에 수록하여 처리한다. 사전에 수록되지 않은 경우는 미등록어 처리로 해결한다.

4. 시스템의 구성 및 특징

제안하는 시스템은 [3]에서 사용한 기존 형태소 분석기에 전처리와 후처리를 추가하여 (그림 1)과 같다. 표제어를 구성하는 전자 사전은 트라이 구조[5]를 사용하며 확장된 조사와 어미 사전을 별도로 사용한다.

(그림 1) 시스템 구성도



(Figure 1) Figure of System

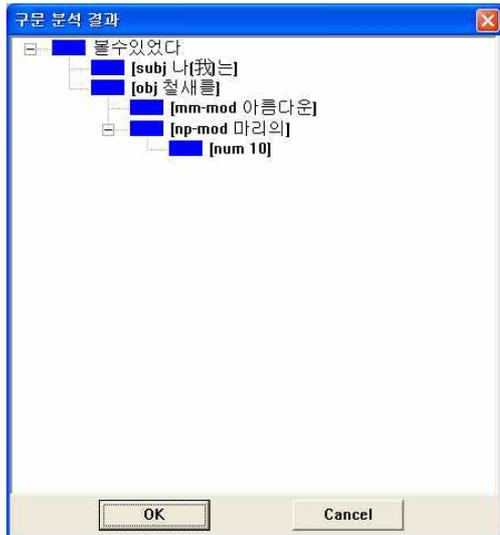
“나(我)는 아름다운 10 마리의 썬새를 볼 수 있었다.”

조사+어미[다]”와 “[접미사]+조사” 형태가 부착된 체언의 미등록어 처리를 하는데 사용된다.

```
int NounProcessing(uchar* rJc)
{
    // 명사류의 접미사 부착
    HasNounSuffix(rJc);
    // 동사파생 접미사 부착
    HasVerbSuffix(rJc);
    // 서술격 조사 '이' 부착
    if ( HasEJosaSuffix(rJc) {
        // 남아있는 rJc에 어미 존재 체크
        VerbProcessing(rJc, NULL);
    }
    // 조사
    HasJosa(rJc);
}
```

제안한 방법으로 분석된 형태소 분석기의 결과를 이용하여 구문 분석을 수행하면 구문 분석용 문법이 간단해짐으로 인해 구문 분석 시간 및 구문 모호성이 감소한다. 본 논문에서 제안한 방법으로 [9]를 사용하여 구문 분석한 결과는 (그림 4)와 같다.

(그림 4) 제안한 형태소 분석 결과를 이용한 파싱



(Figure 4) a parse tree with the proposed morphological analysis

5. 실험 및 평가

성능은 두 방법으로 평가하였다. 첫째는 미등록어 처리와 숫자 처리의 개선으로 기존 시스템에 비해 얼마나 성능이 향상되었는지를 평가하였고 둘째는 제안한 형태소 분석 결과를 이용하여 구문 분석을 직접 실행하였을 때의 성능을 평가하였다. 실험 환경은 <표 1>과 같다.

5.1 기존 시스템과 비교

기존의 형태소 분석 결과와 비교하기 위해서 미등록어와 숫자를 포함한 실험 문장(<표 1> 참조)으로 수행하였다. 미등록어 분석은 체언과 용언의 미지어 추정과 전체적인 형태소 분석기의 정확도 증가에 대해 평가하였고 형성 규칙과 제약 조건을 통한 다 어절 수사 처리 부분은 형태소 분석 후보의 감소에 따른 결과를 통해 정확도를 평가하였다.

<표 1> 실험 환경

OS	Windows 7
CPU	Intel Core i3 CPU 2.4Ghz
Main Memory	4.0GB
Language	Visual C++ 8.0
Dictionary size	120,000 words (nouns 90,000)
Memory requirements	5776KB (first), 5784 (max.)
size of EXE file	152KB
Experimental statement [6]	1,600 sentences consist of 23,920 words

<Table 1> Experimental environment

일반적인 방법의 형태소 분석과 비교해 보았을 때 체언에 대한 미지어 분석은 4.8%, 용언에 대한 미지어 분석은 3.3%, 수사 분석은 3.9% 정확도가 향상되었다.

<표 2> 미지어와 수사처리의 성능 평가

	Frequency	Hit	Fail	Err	Hit Ratio
Unknown Words	체인 : 428	413	15	0	96.5
	용언 : 157	147	7	3	93.6
Numbers	172	168	3	1	97.7

<Table 2> Performance evaluation of processing Unknown-words and Numbers

5.2 파싱에 대한 성능 평가

일반적인 형태소 분석 결과와 본 논문에서 제안한 방법으로 형태소 분석한 결과를 이용하여 파싱을 수행했을 때 생기는 구문 트리의 수를 측정하여 성능을 평가하였다. 본 논문에서는 “확장된 문형을 이용하는 내포문 분할 기반 한국어 구문 분석”[9]에서 제안한 파서를 사용하여 1,600문장을 실험하였다. [3]에서 제안한 형태소 분석 결과를 이용하면 구문 트리는 평균 68.43개가 생성되었지만 제안한 방법은 평균 18.21개의 구문 트리가 생성되어 평균 73.4%가 감소되었다. 또한 [3]의 형태소 분석 결과를 이용해서 구문 분석을 수행할 때는 평균 2.06초 소요되던 것이 제안한 방법을 이용하면 <표 3>과 같이 1.09초로 평균 52.9% 감소하였다.

<표 3> 구문 분석의 성능 평가

Experimental sentence	Resultse (general system)	Results (proposed system)
parse tree	68.43	18.21
Time(sec)	2.06	1.09

<Table 3> Performance evaluation on a parser

6. 결론 및 향후 연구 과제

본 논문에서는 기능어와 분석 단어의 전후 정보를 이용하여 미등록어를 추정하였고 수사의 합성 규칙 및 제약 조건을 이용한 다어절 수사 처리 방법과 문장 속에 나타나는 숫자나 심볼 등을 효과적으로 처리하기 위한 방안을 제시하였으며 복합 명사 및 복합 동사를 형태소 분석 단계에서 처리하였다. 또한, 형태소 분석 후처리

과정에서 복합 명사나 다어절 수사 등에서 의미적으로 묶일 수 있는 것들은 최장일치로 묶어 하나의 구문 단위로 처리함으로써 구문 분석의 부담을 줄여 신속하고 효율적인 구문 분석이 가능하도록 하였다. 제안한 방법은 일반 형태소 분석 결과를 이용할 때보다 구문 트리는 평균 73.4% 감소되었고 분석 소요 시간도 평균 52.9%로 줄어들었음을 알 수 있었다.

향후 연구 과제로는 조사가 생략된 형태의 복합 명사를 합성하거나 미등록어를 추정할 때 문형 정보를 도입하는 방안에 대한 연구가 필요하다.

References

[1] Kang, Seungsik, “Morphological Characteristic and Morphological Analysing technique of Korean,” Journal of KIISE, vol.12, no.8, pp. 47-59, 1994.

[2] Kim, kicheol, “Morphological Analysis-driven Processing of Compound Verbals for Effective Korean Analysis,” Ph.D. thesis, Jeonbuk National University, 1995.

[3] Hwang, Igyu, “Morphological and syntactic ambiguity reduction using syntactic morpheme,” Ph.D. thesis, Jeonbuk National University, 2001.

[4] Sim, kwangsub, “Syllable-based POS Tagging without Korean Morphological Analysis,” The Korean Society for Cognitive Science, Journal of Cognitive Science, vol.22, no.3, pp. 327-345, 2011. 9.

[5] Kim, sangwoo, Seo jeongyeon, “Light Weight Korean Morphological Analysis Using Left-longest-match-preference model and Hidden Markov Model,” The Korean Society for Cognitive Science, Journal of Cognitive Science, vol.24, no.2, pp. 95-109, 2013. 6.

[6] Koh, seunghui, “Construction of syntactico-semantic language resources for the efficient processing of noun sequential structures in Korean.,” Foreign Language University Ph.D. thesis, 2008.

[7] Kim, cheolsu, "An Efficient Electronic Dictionary Structure for the Korean Morphological Analysis Environment," Ph.D. thesis, Jeonbuk National University, 1998.

[8] KIBS : Korean Information Base System, <http://kibs.kaist.ac.kr/kibs>.

[9] Lee, Hyeonyeong, "Syntactic Analysis Based on Embedded Clausal Segmentation Using Extended Sentence Patterns Information", Ph.D. thesis, Jeonbuk National University, 2008.



강 병 도

1986년 서울대학교 계산통계학과 (이학사)
 1988년 서울대학교 대학원 계산학과(이학석사)
 1995년 서울대학교 대학원 계산학과(이학박사)

1998~현재 대구대학교 컴퓨터·IT공학부 교수
 관심분야 : 소프트웨어 공학(Software Engineering), 자연어 처리(NLP), 소프트웨어 아키텍처(Software Architecture) 등



이 현 영

1991년 전북대학교 전산통계학과 (이학사)
 1996년 전북대학교 대학원 전산통계학과(이학석사)
 2008년 전북대학교 대학원 전산통계학과(이학박사)

2014년~2016년: 우석대학 게임콘텐츠학과 조교수
 2016년-현재: 군산대학교 산학협력중점교수
 관심분야 : 자연어처리(NLP), 게임공학(Game Engineering), 빅데이터분석(BigData Analytics), 등.



양 승 원

1985년 전북대학교 전산통계학과 (이학사)
 1987년 전북대학교 대학원 전산통계학과(이학석사)
 1995년 전북대학교 대학원 전산통계학과(이학박사)

1994~현재 우석대학 게임콘텐츠학과 교수
 관심분야 : 자연어처리(NLP), 게임공학(Game Engineering), 데이터베이스(DB), ICT방재(ICT Disaster Prevention) 등



이 중 석

1988년 서울대학교 계산통계학과 (이학사)
 1990년 서울대학교 대학원 계산학과(이학석사)
 2001년 서울대학교 대학원 컴퓨터공학과(공학박사)

1993~현재 우석대학 게임콘텐츠학과 교수
 관심분야 : 소프트웨어 공학(Software Engineering), 자연어 처리(NLP), 객체지향 분석 및 설계(Object-Oriented Analysis and Design) 등