論 文

통계정보에 기반을 둔 한국어 어휘중의성해소

正會員 李 夏 圭* 正會員 金 榮 澤*

Korean Lexical Disambiguation Based on Statistical Information

Ha Gyu Lee*, Young Taek Kim* *Regular Members*

요 약

어휘중의성 해소는 음성 인식/생성, 정보 검색, 말뭉치 태깅 등 자연언어 처리에서 가장 기초가 되는 분야 중의 하나이다. 본 논문은 말뭉치로부터 추출된 통계정보를 이용하는 한국어 어휘중의성해소 기법에 대해 기술한다. 이 기법에서는 좀더 정밀한 중의성해소를 위해 품사태그 대신 형태소분석 결과에 해당하는 토큰태그를 사용하고 있다. 본 논문에서 제안한 어휘선택함수는 어미나 조사의 호응 관계등 한국어의 어휘적 특성을 잘 반영하기 때문에 상당히 높은 정확성을 보여준다. 그리고 활용 분야에 적합하게 사용될 수 있도록 유일선택 방식과 다중선택 방식이라는 두 가지 중의성해소 방식을 지원하고 있다.

ABSTRACT

Lexical disambiguation is one of the most basic areas in natural language processing such as speech recognition/synthesis, information retrieval, corpus tagging, etc. This paper describes a Korean lexical disambiguation mechanism where the disambiguation is performed on the basis of the statistical information collected from corpora. In this mechanism, the token tags corresponding to the results of the morphological analysis are used instead of part of speech tags for the purpose of detail disambiguation. The lexical selection function proposed shows considerably high accuracy, since the lexical characteristics of Korean such as concordance of endings or postpositions are well reflected in it. Two disambiguation methods, a unique selection method and a multiple selection method, are provided so that they can be properly according to the application areas.

Ⅰ. 서 론

영어와 같은 서구어에서 태깅(tagging)이라는 용어는 일반적으로 품사태깅(part of speech tagging)을 지칭하며 한 어절에 대해 세분화된 품사 즉 품사태그(part of speech tag)를 부가하여 어휘중의성(lexical ambiguity)을 해소하는 것을 의미한다[1, 2]. 예를 들어 'table'이라는 어절이 동사로 사용되었으

* 서울大學校 컴퓨터工學科
Dept. of Computer Eng., Seoul National Univ.
論文番號 : 94-26

면 'table/VB'와 같이 'VB(verb no inflection)'라는 태그를 부가하여 이것이 명사가 아니라 동사 원형이라고 결정하여 줌으로써 어휘중의성을 해소한다.

한국어에서도 이와 같이 품사태그를 이용한 방법이 시도된 적이 있었는데[3], 이 방법으로는 약 400개의 많은 태그를 사용함에도 불구하고 위형을 결정할 수 없는 경우가 종종 있다는 문제점을 지니고 있다. 예를 들어 '손'이라는 어절에 대해 '손/Ya'와 같이 'Ya(용언+어말어미)'라는 태그가 부가되었을 때, 이 어절의 명사 '손'이 아니라는 것을 알 수 있지만 '주로'인지 '손르'인지는 결정할 수가 없다. 원형의 결정이 대강의 가장 중요한 기능 중의 하나라는 점을 감안할 때, 이와 같은 방법은 한국어 태깅에서는 적합하지 못함을 알 수 있다. 따라서 한국어 태깅 결과는 형태소분석 결과에 해당하는 토큰(token)수준의 정보를 유지해야 할 필요성이 있다. 예를 들어 '손'이라는 어절이 '주로'로 사용되었다면 '결과:V르:F]'와 같은 형태로 태그를 부가하여 '손'이라는 동사(V)와 '르'이라는 어말어미(F)로 구성되었다는 것을 나타내야만 어휘중의성을 제대로 해소할 수 있다. 그리고 한국어에서 품사태그를 사용할 때 또 다른 어려운 점은 상당히 큰 말뭉치(corpus)가 요구된다는 것이다. 왜냐하면 한국어는 어절 구성이 형어등에 비해 훨씬 복잡하기 때문에 이를 제대로 표현하려면 상당히 많은 태그의 집합이 필요한데, 통계적 방법을 이용할 때 태그의 수가 많으면 많을수록 그만큼의 더 큰 말뭉치를 이용해야만 하고 통계적가 별로 없는 의미있는 통계정보를 추출할 수 있기 때문이다.

이상의 문제점을 고려하여 본 논문에서는 한국어 태깅에서 품사태그 대신에 토큰태그(token tag)를 이용할 것을 제안한다. 그리고 한국어는 복잡한 어절 구성으로 인해 영어에서와 같이 형태소분석을 하지 않고 태깅을 하기에는 어려운 점이 많기 때문에 형태소분석기가 태깅에 필수적인 요소이다. 따라서 본 논문에서 제안한 한국어 태깅 시스템은 형태소분석기와 형태소분석 결과로부터 어휘중의성을 해소하는 어휘중의성해소기(lexical disambiguator)로 구성된다.

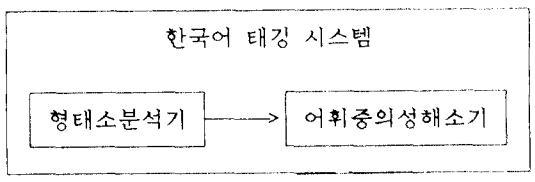

그림 1. 한국어 태깅 시스템의 구성

이제까지의 어휘중의성해소 방법을 살펴보면 크게 어휘중의성해소 규칙을 사용하는 규칙 접근방법과 말뭉치로부터 추출된 통계정보에 기반을 둔 통계적 접근 방법으로 구분할 수 있다[2]. 규칙 접근 방법은 어휘중의성해소 규칙의 구축이 어렵고 정확도가 낮으며 다른 처리영역에서의 적응성이 떨어지기 때문에 현재는 부분적인 특정 문제의 해결 이외에는 거의 사용되지 않는다. 이에 반해 통계적 접근 방법은 정확도가 높을 뿐만 아니라 처리영역이 바뀌어도 일반적인 어휘에 대해서도 그대로 적용될 수 있고 필요하다면 새로운 처리영역의 말뭉치로부터 어렵지 않게 통계 정보를 추출하여 이용할 수 있기 때문에 적응성의 면에서 뛰어난 장점을 지니고 있다. 이러한 이유로 인해 본 태깅 시스템에서는 통계적 접근 방법을 취하고 있다.

앞의 그림 1에서 알 수 있듯이 어휘중의성해소 단계의 입력은 형태소분석기의 결과인 어절들의 리스트 $<w_1 \ldots w_i \ldots w_n>$이다. (이하 본 논문에서 '<...>'는 그 요소들의 순서열을 나타내는 리스트 기호) 그리고 한 어절 w_i는 어휘중의성을 지닐 수 있으므로 토큰들의 집합 $w_i = \{t(i, 1), \ldots, t(i, j), \ldots, t(i, m)\}$가 된다. (여기서 t(i, j)는 i번째 어절의 j번째 토큰) 예를 들어 '나는 간다'라는 입력 문장이 주어지면 '간다'는 두번째 어절이고 형태소분석 결과는 '[가:V 다:F]'와 '[간:V 다:F]'의 두 토큰이 되므로 $w_2 = \{[가:V 다:F], [간:V 다:F]\}$가 된다.(이러한 토큰의 표현에 대해서는 2장 참고) 그러면, 어떤 어절 w_i의 토큰 수를 s_i라 할 때, 한 문장에 대해 총 $s_1 * \ldots * s_i * \ldots * s_n$ 가지의 선택 가능한 토큰리스트(어절당 하나씩 추출된 토큰들의 리스트)가 존재하는데, 어휘중의성해소기의 역할은 이 중에서 최적의 토큰리스트를 선택하는 것으로 볼 수 있다. 물론 필요에 따라 한 어절에 대해 반드시 하나의 토큰만 선택할 수도 있고 가능성이 높은 복수 개의 토큰을 선택할 수도 있다.(6장 참고) 이 선택 과정에서 가장 중요한 점은 최적의 토큰리스트에 대해 최대화될 수 있는 어휘선택함수 f를 어떤 식으로 정의하는가 하는 것이다. 즉,

(1) $f(<t(1, x_1) \ldots t(i, x_i) \ldots t(n, x_n)>) = ?$
(여기서 $t(i, x_i)$는 i번째 어절의 임의의 한 토큰)

통계적 접근 방법에서는 말뭉치로부터 추출된 통계

적인 정보를 이용하여 이 어휘선택함수를 정의하게 되므로 어떠한 말뭉치로부터 어떤 통계정보를 추출하여 이용하느냐가 주된 관심사이다.

이상의 설명에 보충하여 하나 분명히 밝혀 둘 것은 어휘중의해소 단계의 역할이 형태소분석기에 의해 한 어절이 다중 토큰으로 분석되었을 때 그 중의성을 해소하는 것이지 결코 어의중의성해소(wordsense disambiguation)[4]는 시도하지 않는다는 점이다. 이 말은 '시장'이란 어휘가 의미하는 바가 '市長'인지 아니면 '市場'인지를 결정하는 것과 같은 어의 문제는 고려되지 않으며, 어휘중의성해소기가 형태소분석기와 밀접한 관계를 가지고 있고, 따라서 어휘중의성해소기의 정확도에 대한 판단 기준도 형태소분석기와 관련되어 고려되어야 함을 뜻한다.

Ⅱ. 형태소분석기

어휘중의성해소기가 형태소분석기와 밀접한 관계를 맺고 있으므로 2장에서는 본 태깅 시스템에서 사용한 한국어 형태소분석기[5]의 특성 및 처리능력에 대해 간단히 설명하기로 한다.

이 형태소분석기는 비교적 간단한 주범주 체계를 유지하고 있다. 주범주에는 N(명사 : 수사, 대명사 포함), V(동사 : 형용사 포함), ADV(부사), DET(관형사), EXCL(감탄사), ASC(비한국어 어절 : 문장부호 포함), NC(복합명사), NCK(추정명사))의 8가지가 있으며 이 주범주보다 더 정밀한 분류는 종범주에 의해 이루어진다. 예를 들어 동사 즉 'V'로 분석된 어휘에 대한 자동사, 타동사, 형용사 등의 구분은 종범주에 의해 이루어진다. 이러한 종범주의 정확한 결정도 어휘중의성 해소의 한 부분이 될 수 있지만 토큰의 구성 자체에는 영향을 주지 않으므로 본 태깅 시스템에서는 종범주 부분은 고려하지 않으며 앞으로 토큰의 표기에서 종범주에 해당하는 부분은 모두 생략하고 정확도를 언급할 때도 이 부분은 무시하기로 한다.

(2) 아름다운 : [아름답 : V/AJ-ㄴ : F]
(AJ : 형용사에 해당하는 종범주, F : 어미 혹은 어미 + 조사)

한 어절이 어휘중의성을 가지고 있으면 다중 토큰으로 분석되고, 조사나 어미의 이형태는 그대로 출력되며 서술격조사, 명사화어미 및 선어말어미는 분리되지만 나머지 조사나 어미가 복합된 형태로 사용되었으면 결합된 형태로 그대로 출력된다. 결합된 형태로 조사나 어미를 출력하는 것이 기계번역등에서 파싱(parsing)과 같은 차후 처리단계에서 문제가 될 수도 있지만 어휘중의성해소에서는 그대로 사용할 수도 있고 또한 필요하다면 간단한 분리표를 사용하여 대체로 유일하게 분리가 가능하므로 실제로는 거의 문제가 되지 않는다. 그리고 '들', '하' 등 비교적 빈도가 높은 접미사는 분리되며, 결합된 형태로 사용된 보조용언의 분석 기능도 지니고 있다.

(3) 한 : [한 : N][하 : V-ㄴ : F][한 : DET]
했다고는 : [하 : V-ㅆ : PE-다고는 : F]
학교가 : [학교 : N-가 : P]
사람들이 : [사람 : N-들 : SF-이 : P]
집이 : [집 : N-이 : P]
공부하면 : [공부 : N-하 : SF-면 : F]
학교임을 : [학교 : N-이 : CP-ㅁ : F-을 : P]
해주기 : [하 : V-아 : F-주 : VX-기 : F]
(P : 서술격조사를 제외한 조사, CP : 서술격조사, PF : 선어말어미, SF : 접미사, VX : 보조용언)

뿐만 아니라, 위에서 언급된 주범주 체계에서 알 수 있듯이 이 형태소분석기는 복합명사 및 미등록어 추정 기능도 지니고 있으며 이는 비한국어가 포함된 어절에 대해서도 마찬가지이다.

(4) 국민학교가 : [국민학교 : NC-가 : P]
차세대전투기를 : [차세대전투기 : NCK-를 : P]
OPEN하면 : [OPEN : ASC-하 : SF-면 : F]
3개의 : [3 : ASC-개 : N-의 : P][3 : ASC-개의 : N]

현재 형태소분석기에 사용되는 사전은 약 11만 표제어를 지니고 있으며 '하다', '적'등 빈도가 높은 접미사로 파생된 어휘는 표제어에 결합정보로 기술되므로 실제 어휘 수는 약 15만개가 된다. 다중 토큰으로 분석된 어절의 경우 그 문맥에 맞는 토큰이 포함되어 있으면 올바르게 분석되었다고 볼 때, 이 형태소 분석기는 평균 99.36%의 정확도를 지니고 있으며 전체 어절에 대해 다중 분석된 어절의 비율은 문장부호를 별도의 어절로 보았을 때 평균 17% 정도가 된다. 여기서 분명히 할 것은 엄밀한 의미에서 '어절'이라 함은 띄어쓰기의 단위를 지칭하지만 본 논문의 어휘중의성해소 단계에서는 문장부호를 별도의 어절로 취급하므로 앞으로 문장부호는 별도의 어절로 생각

하기로 한다. 한편, 다중 분석된 어절의 경우 약 99.85% 정도가 정확하며 한 어절당 평균 2.3개의 토큰이 생성된다. 그리고 중의성이 없이 하나로 분석된 어절에 대해서는 평균 99.26%의 정확도를 지니고 있다.

Ⅲ. 토큰 구성형식

어휘중의성해소 단계에서는 이 단계에 적합하게 형태소분석 단계에서 출력된 토큰을 재구성하여 어휘중의성해소에 사용한다. 3장에서는 이러한 어휘중의성해소 단계에 사용되는 토큰의 구성형식을 설명하는데 이는 어디까지나 중의성해소 단계에서 내부적으로 사용되는 것으로 4장에서 설명될 통계정보 추출까지는 이 토큰 구성형식이 적용되지만 어휘중의성해소 단계의 출력은 형태소분석 단계에서 출력하는 토큰의 형식을 그대로 유지한다. 먼저 토큰의 구성을 설명한 다음 어휘중의성해소 단계에서 사용하는 품사 체계에 대해 설명한다.

어휘중의성해소 과정에서는 한 토큰을 머리와 꼬리로 구분하여 사용한다. 머리는 한 토큰의 중심이 되는 요소로 이루어지며 대체로 내용어 부분에 해당한다고 볼 수 있다. 머리에는 2장에서 설명된 형태소분석 결과 중에서 주범주를 할당받는 토큰의 구성요소와 서술격조사, 명사화어미, 접미사가 포함되는데, 접미사 중에서 중의성해소에 별 영향을 주지 못하는 '들', '등', '뿐'등은 제외된다. 그리고 결합된 형태로 사용된 보조용언이 나타난 토큰의 경우 보조용언과 그 앞에 온 어말어미도 머리에 포함시킨다. 꼬리는 한 토큰에서 주로 문법적 기능을 나타내는 부분으로 이에는 어절 끝에 오는 어말어미와 조사가 포함된다. 이와 같은 토큰의 구분에 따르면 형태소분석기가 출력하는 토큰의 구성요소 중에는 머리와 꼬리 어느 부분에도 속하지 않는 것이 있는데, 일부 접미사와 선어말어미 및 토큰의 중간에 오는 조사가 이에 해당하며 이러한 부분은 어휘중의성해소에 별 도움이 되지 못하므로 중의성해소 단계에서는 무시한다.

이상과 같은 머리와 꼬리의 구분은 이웃한 두 토큰 사이의 호응 관계를 포착하기 위한 것인데 머리에 서술격조사와 명사화어미를 포함한 이유는 두 토큰 사이의 호응 관계가 이들이 나타난 경우와 그렇지 않는 경우에 다르기 때문이다. 예를 들어 서술격조사가 붙은 '우등생이었다'와 같은 경우에는 동사를 수식하는 '학창시절에'와 같은 부사어의 수식을 받을 수 있지만 서술격조사가 없는 '우등생은'과 같은 경우에는 일반적인 부사어의 수식을 받을 수 없다.(이러한 호응 관계에 대해서는 5.1 참고) 다음은 머리와 꼬리의 구분에 대한 예를 보여준다.

(5) 공부하면 : [공부 : V 하 : SF 면 : F] -> 공부하-면
했다고는 : [하 : V 었 : PF 다고는 : F] -> 하-다고는
학교에서부터이지만 : [학교 : N 에서부터 : P 이 : CP 지만 : F] -> 학교에 지만
사람들만이 : [사람 : N 들 : SF 만이 : P] -> 사람 만이
사람임이 : [사람 : N 이 : CP ㅁ : F : P] -> 사람임 이
해주기 : [하 : V 어 : F 주 : VX 기 : F] -> 하어주기 null

어휘중의성해소 단계에서는 머리 부분에만 품사를 할당하여 사용한다. 물론 꼬리 부분도 중의성해소에 큰 영향을 미치므로 품사태그를 이용할 경우라면 꼬리 부분도 품사 설정에 포함되어야 하지만 본 대상 시스템에서는 토큰내 1개 사용하고 꼬리 부분의 영향은 다른 유형의 통계정보로써 포착하고자 시도했기 때문에(5.1 참고) 중복을 피하고 희소 통계치를 줄이기 위해 꼬리 부분에는 품사를 할당하지 않는다. 이외 같이 머리에만 품사를 할당하여도 이론적으로 가능한 경우는 30가지가 넘는데 4장에서 설명될 말뭉치에서 실제로 나타난 품사는 다음 16가지밖에 없었다.

(6) N, V, ADV, DET, EXCL : 형태소분석기의 주범주와 동일
NV : N에 서술격조사가 붙은 경우(예 : 학생이 다)
NVN : NV에 명사화어미가 붙은 경우(예 : 학생임 이)
VN : V에 명사화어미가 붙은 경우(예 : 사랑하기 가)
VV : 보조용언이 결합된 경우(예 : 하어주-는)
VVN : VV에 명사화어미가 붙은 경우(예 : 하어주기-도)
NN : 수사+N의 경우(예 : 3개 를)
NNV : NN에 서술격조사가 붙은 경우(예 : 3개이-고)
NNVN : NNV에 명사화어미가 붙은 경우(예 : 3개임 이)
P : 문장부호
SB : 문장시작
SE : 문장끝

여기서 주목할 것은 문장의 시작과 끝도 중의성해소

에는 의미가 있는 부분이기 때문에 중의성해소 단계에서 사용하는 것이 바람직하므로 특별한 품사를 할당받았고, 이에 따라 4장에서 설명될 말뭉치에서 추출된 통계정보와 5장에서 설명될 어휘선택함수의 통계 수식에 문장시작과 문장끝을 의미하는 두 개의 토큰 t(0, 1)과 t(n+1, 1)이 있다는 가정을 한다.

Ⅳ. 말뭉치와 통계정보

4장에서는 통계정보 추출에 사용된 말뭉치의 상태 및 특성과 이 말뭉치로부터 추출된 통계정보에 대해 설명한다.

통계정보 추출에 사용된 말뭉치는 전산학에 관련된 일반 서적 및 과제 수행 보고서 10권으로부터 수집된 약 33만 어절로 구성되어 있다. 이 말뭉치의 구축 과정을 간단히 살펴보면 먼저 사람이 띄어쓰기 및 철자 교정을 마친 다음, 형태소분석을 하였다. 그리고 형태소분석 결과 중에서 빈도가 비교적 높은 어절의 오류 및 전형적인 오류를 바로잡은 후, 어휘중의성을 제거하였다. 중의성 제거 과정에서 처음 핵심자료 5만 어절은 전부 사람의 수작업으로 이루어졌으며, 나머지 28만 어절은 6.2에서 설명될 다중선택 방식을 되먹임 방법으로 적용하여 구축되었다. 즉, 중의성해소 프로그램을 이용하여 일차적으로 어느 정도의 중의성을 자동적으로 제거한 후, 미해결된 부분만 사람이 중의성을 제거하고 다시 이들 통계정보를 기존의 통계정보에 병합하여 다음 자료의 중의성 제거에 이용하는 방법을 반복적으로 적용하였다. 현재 이 말뭉치는 약간의 띄어쓰기 및 철자 오류, 형태소분석기의 오류, 중의성해소 프로그램의 오류, 사람의 선택 오류 등을 포함하여 전체적으로 보아 약 1%의 오류를 포함하고 있는데 실험 결과를 검토해 본 결과 이 정도의 잡음도는 중의성해소에 별 영향을 주지 않는 것으로 밝혀졌다.

이 말뭉치의 통계적 특성을 살펴보면, 2장에서 설명된 형태소분석기를 사용하였을 때, 각 어절은 평균 7.9번 사용되었는데 이를 영어의 DOE(Dept. of Energy) 말뭉치 33만 어절에 대해 조사해 본 결과인 16.5번과 비교해 보면 한국어가 훨씬 다양한 어절 형태로 사용됨을 알 수 있다. 그리고 다중 분석된 어절(전체 어절의 약 17%)의 경우 평균 11.2번 사용되었고 중의성이 없이 분석된 어절은 평균 7.2번 사용되었다. 이로부터 한국어 어절의 경우 중의성을 지닌 어절이 그렇지 않은 어절보다 자주 사용됨을 알 수 있다. 기본형으로 본 서로 다른 어휘의 수는 약 13,000인데 이 크기는 전체 말뭉치 크기의 약 3.9%에 해당한다.

위의 말뭉치에서 추출된 통계정보는 7가지 유형이 있는데 다음은 그 각각에 대해 자료의 크기 및 예를 보여준다.

(7) 토큰별 발생 빈도*: 4,998
(한 588 ([한:N] 24) ([하:V-ㄴ:F] 68) ([한:DET] 496))
품사 빈도*: 16
(N 133774)
(V 69460)
bigram 빈도: 132
(〈N V〉45274)
trigram 빈도*: 743
(〈V N V〉13260)
머리의 빈도*: 15,499
(하:V 4188)
꼬리의 빈도: 383
(라:F 568)
꼬리-머리의 공기(cooccurrence) 빈도*: 47,864
(〈라:F 하:V〉80)

여기서 토큰별 발생 빈도는 다중 분석된 어절에서 각 토큰이 말뭉치에 나타난 빈도를 말하며 bigram과 trigram 빈도는 각각 연속된 두 개 및 세 개의 토큰에 대한 머리의 품사 빈도를 말한다. 꼬리-머리의 공기 빈도는 연속된 두 토큰에서 앞 토큰의 꼬리와 뒤 토큰의 머리가 함께 나타난 빈도로 주어진다. 그리고 꼬리의 빈도와 꼬리-머리의 공기 빈도에서 이형태의 조사와 어미들은 그 대표형 한 가지로만 추출하였다. 이들 통계정보 중에서 '*' 표시가 된 것들만 최종적으로 선정된 어휘선택함수에서 사용되고 나머지는 단지 5.1에서 설명될 배경실험에만 사용됨을 미리 밝혀둔다.

Ⅴ. 어휘선택함수

5장에서는 한국어 태깅에 적합한 어휘선택함수의 선정 과정에 대해 설명한다. 통계적인 접근방법에서 태깅에 적합한 어휘선택함수를 구할 때는 '어떠한 유형의 통계정보를 이용할 것인가' 그리고 각 유형의 통계정보에 대해 '어떤 통계 수식으로 이들 정보를

어휘선택함수에 반영할 것인가'가 매우 중요하다. 본 논문에서는 세가지 유형의 통계정보를 설정하고 각 유형의 통계정보에 대해 몇 가지 통계 수식을 마련하고 각각에 대해 개별적으로 정확도를 측정하는 배경실험을 통해 각 유형별로 나타나는 특성을 이용 조합하여 최종적인 어휘선택함수를 설정하는 방식을 취하였다. 먼저 어휘선택함수 설정을 위해 수행된 배경실험에 대해 살펴보고 다음에 최종적으로 설정된 어휘선택함수에 대해 설명하기로 한다.

5.1 배경실험

본 태깅 시스템에서는 통계정보의 유형으로 한 어절 내에서의 '토큰 확률', 이웃한 어절 사이에서의 '품사 관계' 및 '어미/조사 호응 관계' 세 가지를 설정하였다. 이 세 가지 유형의 통계정보는 그들 사이에 관련성이 비교적 없는 거의 독립적인 것들이다. 이렇게 독립적인 유형의 정보를 설정한 이유는 최종적인 어휘선택함수에 이들 정보를 조합하여 반영할 때 독립적일수록 그 상승 효과가 크기 때문이다. 여기서 앞의 두 가지 유형의 정보는 그 세부적인 적용 방식은 약간씩 차이가 나지만 이미 영어에서도 논의된 바 있고[2, 6, 7], 여기에 어미/조사의 호응 관계를 추가한 이유는 한국어에서 어미와 조사가 풍부한 문법적 정보를 포함하고 있어서 다음에 오는 어휘와 밀접한 관련성을 지니고 있으므로 어휘중의성해소에 아주 유용하기 때문이다.

다음은 배경실험의 방법 및 결과에 대해 통계정보의 유형별로 설명한다. 이에 앞서 먼저 실험에 사용된 검증자료에 대해 살펴보면, 이 검증자료는 4장에서 설명된 통계정보 추출에 사용된 말뭉치에 포함되지 않은 것으로 두 권의 일반적인 전산학 서적으로부터 무작위로 추출되었다. 이 자료는 1,212 문장, 15,443 어절로 이루어져 있으며, 다중 분석된 어절은 2,610개로 전체 어절의 약 16.9%에 해당한다. 그리고 이하 배경실험에서 '정확도'는 한 어절에 대해 항상 하나의 토큰을 선택하는 '유일선택 방식'을 적용했을 때 다중 분석된 어절 중에서 올바르게 선택된 어절의 비율을 말하며, 통계 수식의 계산에서 적용될 통계치가 없거나 0인 경우에는 말뭉치에서 나타난 0이 아닌 최소값을 사용하였다.(유일선택 방식에 대해서는 6.1 참고)

5.1.1 토큰 확률에 대한 실험

'토큰 확률'에 대해서는 다음 (8)의 '방법 T/W' 하나만 실험하였는데 다른 유형의 정보보다 중의성 해소에 효과가 매우 큰 것으로 밝혀졌다. 이 방법에서 어휘선택함수 f는 주어진 토큰리스트에서 각 어절 w_i가 토큰 $t(i, x_i)$인 확률의 곱으로 정의되는데 이 확률은 말뭉치에서 추출된 통계정보를 이용하면 근사적으로 (8)의 'prob T/W'와 같다.(실제로 이 방법에서는 어절 단위로 선택을 수행해도 결과가 같은데, 다른 유형의 통계정보와 조합할 경우를 고려하여 일관성 있게 표현했음) 여기서 한 가지 흥미로운 사실은 다중 분석된 어절 중에서 말뭉치에서 한번 이상 나타난 어절에 대해 적용했을 때는 약 92%가 적용되었으며 이 경우의 정확도는 95.4% 정도로 비교적 높았는데, 이로부터 말뭉치가 커서 적용 가능한 어절의 수가 증가하면 이 방법은 좀더 정확해질 수 있음을 알 수 있다.

(8) 방법 T/W(정확도 : 91.25%)

$$f \doteqdot \prod_{i=1}^{n} \text{Prob T/W}(t(i, x_i)), \text{prob T/W}(t(i, x_i)) \doteqdot \frac{\text{freq}(t(i, x_i))\, w_i}{\text{freq}(w_i)}$$

예) prob T/W([하 : V-L : F]) : 68/588
(이하 예에서 사용된 수치에 대해서는 4장 (7)을 참고)

5.1.2 품사 관계에 대한 실험

'품사 관계'에 대해서는 다음 (9)에 열거한 5가지 방법을 실험하였다.

(9) 방법 X/YZ (정확도 : 51.28%)

$$f \doteqdot \prod_{i=1}^{n} \text{prob-X/YZ}(X|\langle Y\ Z\rangle), \text{prob-X/YZ}(X|\langle Y\ Z\rangle) \doteqdot \frac{\text{freq}(\langle X\ Y\ Z\rangle)}{\text{freq}(\langle Y\ Z\rangle)}$$

예) prob-X/YZ(V|⟨N V⟩) : 13260/45274

방법 Y/Z (정확도 : 44.54%)

$$f \doteqdot \prod_{i=0}^{n} \text{prob-Y/Z}(Y|Z), \text{prob Y/Z}(Y|Z) \doteqdot \frac{\text{freq}(\langle Y\ Z\rangle)}{\text{freq}(Z)}$$

방법 Z/XY (정확도 : 51.56%)

$$f=\prod_{i=1}^{n} \text{prob-Z/XY}(Z|\langle X\ Y\rangle),\ \text{prob-Z/XY}(Z|\langle X\ Y\rangle) \fallingdotseq \frac{\text{freq}(\langle X\ Y\ Z\rangle)}{\text{freq}(\langle X\ Y\rangle)}$$

방법 Z/Y (정확도 : 45.14%)

$$f=\prod_{i=0}^{n} \text{prob-Z/Y}(Z|Y),\ \text{prob-Z/Y}(Z|Y) \fallingdotseq \frac{\text{freq}(\langle Y\ Z\rangle)}{\text{freq}(Y)}$$

방법 MI-XYZ*(정확도 : 68.18%)

$$f=\prod_{i=1}^{n} \text{MI-XYZ*}(X, Y, Z),\ \text{MI-XYZ*}(X, Y, Z) \fallingdotseq \frac{\text{freq}(\langle X\ Y\ Z\rangle)}{\text{freq}(X)*\text{freq}(Y)*\text{freq}(Z)}$$

예) MI-XYZ*(V, N, V) : 13260/(69460*133744*69460)

(X, Y, Z는 각각 토큰 $t(i-1, x_{i-1})$, $t(i, x_i)$, $t(i+1, x_{i+1})$의 품사)

'방법 X/YZ'에서는 오른쪽에 머리의 품사가 'Y', 'Z'인 두 토큰이 왔을 때 현재 위치에 있는 토큰의 머리 품사가 'X'일 확률을 사용하였다. '방법 Y/Z'는 '방법 X/YZ'에서 우측 look-ahead를 하나로 줄인 방법인데, 이 두 방식은 영어에서도 적용된 바 있다.(영어에서는 어절 전체에 품사를 할당하는 품사태그 방법으로 적용한 것이 다름) '방법 Z/XY'와 '방법 Z/Y'는 각각 앞의 두 방법에서 우측 look-ahead 대신 좌측 look-ahead를 사용한 방법인데 이 두 방법을 시도한 이유는 영어는 중심어가 앞에 오지만 한국어는 중심어후행 언어이므로 영어와는 반대 방향의 문맥을 반영할 때의 효과를 검토해 볼 필요가 있었기 때문이다. 그런데 비록 좌측 look-ahead를 사용한 경우가 약간 더 정확했지만 이들 네 방법은 모두 정확도가 매우 낮은 편이다. 특히 look-ahead가 하나인 경우를 살펴보면 무작위로 선택했을 경우보다 별로 나은 바가 없다.(다중 어절의 평균 토큰 수가 2.3이므로 무작위로 선택할 때의 정확도는 43.5%) 이상의 네 방법에서 정확도가 낮은 것에 대해 꼬리 부분의 정보가 반영되지 않았고, 품사가 세분화되지 못했으며 한국어 어순이 비교적 자유롭다는 등의 이유를 생각해 볼 수 있는데 이에 대한 검토는 좀더 연구가 진행되어야 할 것으로 보인다.

이상의 방법이 한국어 태깅에 접합하지 못함에 따라 상호정보[8] 개념을 이용한 '방법 MI-XYZ*'를 시도해봤는데 이 방법은 앞의 방법들에 비해 상당히 높은 정확도를 보였다. 본래의 상호정보 개념을 품사관계에 적용하면 그 수식은 (10)의 'MI-XYZ'와 같이 되며 이것은 세 품사 'X', 'Y', 'Z'가 완전히 독립적이라고 가정할 때 이들이 우연히 이 순서로 이웃하여 나타날 확률에 대해서 이 세 품사가 어떤 문장에서 실제로 이 순서로 이웃하여 나타날 확률의 비를 의미한다. 이러한 확률들 역시 (10)의 근사식과 같이 말뭉치의 통계정보를 이용하여 근사적으로 계산할 수 있는데, 여기서 말뭉치의 크기 'S'는 실제로는 불필요하고 다른 유형의 통계정보에서 계산된 값처럼 0보다 크고 1 이하로 유지하기 위해 이 'S'를 제외한 상호정보 수식 'MI-XYZ*'를 어휘선택함수에 사용하였다.

(10)

$$\text{MI-XYZ}(X, Y, Z) \fallingdotseq \frac{\frac{\text{freq}(\langle X\ Y\ Z\rangle)}{S}}{\frac{\text{freq}(X)}{S}*\frac{\text{freq}(Y)}{S}*\frac{\text{freq}(Z)}{S}} = \frac{\text{freq}(\langle X\ Y\ Z\rangle)*S^2}{\text{freq}(X)*\text{freq}(Y)*\text{freq}(Z)}$$

(S는 통계정보 추출에 사용된 말뭉치의 크기)

5.1.3 어미/조사 호응 관계에 대한 실험

'어미/조사의 호응 관계'에 대해서는 다음 (11)의 두 방법을 실험하였다.

(11) 방법 MI-TH*(정확도 : 72.71%)

$$f=\prod_{i=1}^{n+1} \text{MI-TH*}(T, H),\ \text{MI-TH*}(T, H) \fallingdotseq \frac{\text{freq}(\langle T\ H\rangle)}{\text{freq}(T)*\text{freq}(H)}$$

예) MI-TH*(라 : F, 하 : V) : 80/(568*4188)

방법 T/H (정확도 : 77.07%)

$$f=\prod_{i=1}^{n+1} \text{Prob-T/H}(T|H),\ \text{Prob-TH*}(T|H) \fallingdotseq \frac{\text{freq}(\langle T\ H\rangle)}{\text{freq}(H)}$$

예) Prob-T/H(라 : F하 : V) : 80/4188

(T는 토큰 $t(i-1, x_{i-1})$의 꼬리, H는 토큰 $t(i, x_i)$의 머리)

'방법 MI-TH*'는 연속한 두 토큰에서 앞 토큰의 꼬리와 뒤 토큰의 머리 사이의 상호정보 수식(맙붕치 크기 제외)을 사용한 방법이며, '방법 T/H'에서는 어떤 머리 'H' 바로 앞에 꼬리 'T'가 올 확률을 사용하였다. '방법 MT-TH*'에는 꼬리의 빈도라는 통계치가 하나 더 도입되었음에도 불구하고 이 두 방법 중에서 '방법 T/H'가 약간 더 우수했는데, 맙붕치가 지금보다 많이 확장되었을때도 '방법 T/H'가 더 좋을지는 예측하기 힘드므로 이에 대한 연구는 좀더 진행되어야 할 것으로 보인다.

5.2 제안된 어휘선택함수

본 태깅 시스템에서는 5.1의 배경실험 결과를 바탕으로 각 통계정보의 유형별로 정확도가 가장 좋은 방법을 한 가지씩 골라 조합하여 최종적인 어휘선택함수로 선정하였다. 이는 다음 (12)의 f와 같이 '토큰 확률', '품사 관계', '어미/조사 호응 관계' 각각에서 시도된 '방법 T/W', '방법 MI-XYZ*' 및 '방법 T/H'에서 사용된 개별적인 어휘선택함수의 곱으로 정의된다.

(12) $f(\langle t(0, 1) \ldots t(i, x_i) \ldots t(n+1, 1) \rangle)$

$$= \prod_{i=1}^{n} \text{prob}-\text{T/W}(t(i, x_i)) \times \prod_{i=1}^{n} \text{MI}-\text{XYZ}^*(X, Y, Z) \times \prod_{i=1}^{n-1} \text{Prob}-\text{T/H}(T|H)$$

Ⅵ. 어휘중의성해소 방식

6장에서는 유일선택 방식과 다중선택 방식이라는 두 가지 어휘중의성해소 방식에 대해 설명하고 각각에 대해 실험 결과를 제시하고 고찰하기로 한다.

6.1 유일선택 방식

본 논문에서 '유일선택 방식'이라 함은 한 어절에 대해 항상 하나의 토큰만 선택하는 어휘중의성해소 방식을 말하는데, 이 방식에서의 중의성해소 과정을 가상의 예를 중심으로 설명하기로 한다.

본 태깅 시스템에서 중의성해소 과정은 입력 문장의 왼쪽에서 오른쪽으로 토큰리스트 확장과 토큰리스트 선택을 반복함으로써 진행된다.(오른쪽에서 왼쪽으로 진행해도 결과는 같지만 실시간 처리를 위해 왼쪽에서 오른쪽으로 진행함) 토큰리스트 확장 과정에서는 기존의 토큰리스트 각각에 대해 새로 들어온 어절의 각 토큰을 추가하고 어휘선택함수의 값을 갱신한다.(아래 (13)의 (b)에서 (c)로의 과정 참고) 토큰리스트 선택 과정에서는 방금 확장된 어절보다 두 번째 이전의 어절에 대해 선택이 이루어지는데, 현재 선택하고자 하는 어절의 토큰 수가 하나면 선택이 불필요하다. 예를 들어 이 과정을 구체적으로 살펴보면, 다음 (13)의 (a)에서 두 토큰리스트 〈a1 b1 c1〉과 〈a2 b1 c1〉 중에서 어휘선택함수 값이 높은 〈a2 b1 c1〉만 이 시점에서 선택하고 나머지 〈a1 b1 c1〉을 제거해도 전체적인 결과에는 영향을 미치지 않는다. 왜냐하면 실제로 어휘선택함수의 계산은 이웃한 세 토큰 내에서 이루어지므로 다음에 'D'를 확장하여 어휘선택함수 값을 갱신할 때 'A'의 각 토큰은 더 이상 영향을 주지 못하고, 'B', 'C' 부분이 동일하기 때문이다.

(13) A = {a1, a2}, B = {b1}, C = {c1, c2}, D = {d1, d2, d3}일 때

(a) C까지의 확장 결과
- 〈a1 b1 c1〉: 0.01
- 〈a2 b1 c1〉: 0.06
- 〈a1 b1 c2〉: 0.10
- 〈a2 b1 c2〉: 0.03

(b) A를 선택한 결과
- 〈a2 b1 c1〉: 0.06
- 〈a1 b1 c2〉: 0.10

(c) D의 확장 결과
- 〈a2 b1 c1 d1〉: 0.03
- 〈a1 b1 c2 d1〉: 0.07
- 〈a2 b1 c1 d2〉: 0.05
- 〈a1 b1 c2 d2〉: 0.06
- 〈a2 b1 c1 d3〉: 0.02
- 〈a1 b1 c2 d3〉: 0.01

이상의 선택 과정에서 알 수 있듯이, 이 유일선택 방식은 모든 선택 가능한 토큰리스트 중에서 어휘선택함수 값을 최대화시키는 토큰리스트를 입력 문장의 길이에 대해 선형적인 시간 내에 구할 수 있다. 5.1에서 언급한 검증자료에 대해 (12)의 어휘선택함수를 유일선택 방식으로 적용한 결과 96.33%의 비교적 높은 정확도를 보였다.

6.2 다중선택 방식

태깅의 활용 분야 중에서 한 어절에 대해 반드시 하나의 토큰을 선택해야 하는 경우에는 유일선택 방식이 적합하지만 이 방식의 정확도 96.33%로는 좀더

정확한 어휘중의성해소를 요하는 말뭉치 태깅이나 기계번역등에서 파싱 전에 어휘중의성을 줄이는 것과 같은 분야에는 그대로 사용하기가 어렵다. 따라서 비록 어휘중의성을 완전히 제거하지는 못하지만 태깅에 의한 오류를 가능한 한 적게 하는 어휘중의성해소 방식이 요구되는 경우가 많다. 다음은 이러한 필요성에 적합하도록 한 어절에 대해 가능성이 높은 복수 개의 토큰을 선택하는 '다중선택 방식'에 대해 설명한다.

다중선택 방식은 대체로 유일선택 방식과 유사하게 진행되는데, 그 차이점은 유일선택 방식에서는 어휘선택함수 값을 최대화시키는 하나의 토큰리스트를 선택함으로써 토큰의 선택이 이루어지지만 다중선택 방식에서는 토큰의 선택이 6.1의 토큰리스트 선택 과정 전에 이루어진다. 이 토큰 선택 과정에서는 현재 선택하고자 하는 어절에서 지금까지의 어휘선택함수 값이 최대인 토큰리스트와의 값의 비가 어떤 선택비율 이상인 토큰리스트들에 속한 토큰을 모두 선택하게 된다. 예를 들어, (13)의 (a)에서 〈a1 b1 c2〉가 어휘선택함수 값 0.10으로 최대인 토큰리스트이므로 선택비율이 1이면 〈a1 b1 c2〉만 '0.10/0.10 ≥ 1'이므로 여기서 'a1'이 어절 'A'의 토큰에서 선택된다. 만약 선택비율이 0.5라면 〈a2 b1 c1〉도 '0.06/0.10 ≥ 0.5'로 선택비율 이상이므로 'a2'도 선택되어 결과적으로 'A'에 대해서는 'a1'과 'a2' 두 토큰이 선택된다.

이상과 같이 다중선택 방식에서는 선택비율을 조정함으로써 선택의 정도와 정확도 사이에 취사 선택할 수 있는데, 이 방식에서 유지되는 토큰리스트의 수는 유일선택 방식과 동일하므로 역시 입력 문장의 길이에 대해 선형적인 시간 내에 중의성해소가 가능하다. 다음 〈표 1〉은 5.1에서 언급한 검증자료에 대해 (12)의 어휘선택함수를 다중선택 방식으로 적용했을 때 선택비율에 따른 실험 결과를 보여준다. 이 표에서 정확도는 선택된 토큰들 중 하나라도 문맥에 맞는 것이 있을때 올바르다고 할 때 다중 분석된 어절 중에서 올바르게 선택된 어절의 비율을 말하며 미해결 어절은 다중 분석된 어절에서 두 개 이상의 토큰이 선택된 어절의 비율을 나타낸다. 그리고 어절 당 토큰 수는 다중 분석된 어절에서 중의성해소가 끝나고 남은 어절당 평균 토큰 수를 나타낸다.

선택비율이 감소함에 따라 당연히 정확도가 증가하지만 미해결 어절과 어절당 토큰 수도 이에 따라 증가한다. 선택비율이 0.001인 경우에 대해 좀더 고찰해보면, 다중 분석된 어절에서 어휘중의성해소기의 오류는 0.56% 정도인데 다중 분석된 어절의 비율이 전체 어절의 약 17%이므로 이 오류는 전체 어절에 대해서는 약 0.10%에 해당한다. 이는 1,000개 어절당 평균 1개 꼴로 틀리는 것을 말한다. 그리고 이 때 미해결 어절의 비율은 다중 분석된 어절에 대해서는 26.6%, 전체 어절에 대해서는 약 4.5%가 된다. 이것을 말뭉치 태깅 입장에서 보면 1,000개 어절 중에서 45개 정도만 사람이 개입하면 됨을 의미한다. 어절당 토큰 수의 결과는 기계번역등에서 파싱 전에 어휘중의성을 줄이는데 이 방식을 사용한다면 다중 분석된 어절의 토큰 수를 평균 2.3개에서 1.28개로 줄일 수 있음을 뜻한다.

표 1. 다중선택 방식의 실험 결과

선택비율	정확도(%)	미해결 어절(%)	어절당 토큰 수
1	96.89	2.0	1.02
0.1	98.47	8.4	1.08
0.01	99.16	14.3	1.16
0.001	99.44	26.6	1.28

Ⅶ. 결 론

본 논문은 말뭉치에서 추출된 통계정보를 이용하는 한국어 어휘중의성해소 기법에 대해 기술하고 있다. 이를 기존의 기법과 비교해 보면, 품사태그 대신 토큰태그를 사용함으로써 좀더 정밀한 태깅이 가능하고, 제안된 어휘선택함수가 한국어의 어휘 특성을 잘 반영하고 있기 때문에 정확도가 높다는 특징을 가지고 있다. 그리고 유일선택 방식과 다중선택 방식을 지원하므로 활용 분야에 따라 적합하게 사용될 수 있는 장점도 지니고 있다.

이러한 어휘중의성해소기를 포함하는 태깅 시스템은 음성 인식/생성, 정보 검색, 말뭉치 태깅, 기계번역 등 자연언어 처리의 여러 분야에서 어휘중의성을 해소하는데 유용하게 활용될 수 있다. 본 논문의 어휘선택함수를 유일선택 방식으로 적용하면 음성 인식/생성, 정보 검색 등에서 비교적 높은 정확도로 한 어절에 대해 항상 하나의 토큰을 선택할 수 있다. 그리고 다중선택 방식을 사용하면 반자동으로 말뭉치 태깅을 할 때 사람의 노력을 크게 줄일 수 있으며, 기계번역등에서는 파싱 전에 많은 어휘중의성을 높은 정확도를 갖고 제거할 수 있다.

앞으로 어휘중의성해소의 정확도 및 적응성을 높이기 위해서는 말뭉치를 좀더 확장한 필요가 있으며, 또한 본 논문에서 제안한 어휘선택함수가 다른 처리영역이나 일반적인 처리영역에 적용되었을 때 그 타당성에 대한 검증이 필요하다.

부록 : 어휘중의성해소 결과

입력 문장

이 줄 앞에 한 행을 추가하고, 전의 커서 위치로 다시 이동하시오.

형태소분석 결과

이 : [이 : N][이 : DET][이 : EXCL]
줄 : [줄 : N][주 : V-ㄹ : F][줄 : V-ㄹ : F]
앞에 : [앞 : N-에 : P]
한 : [한 : N][하 : V-ㄴ : F][한 : DET]
행을 : [행 : N-을 : P]
추가하고 : [추가 : N-하 : SF-고 : F][추가 : N-하고 : P]
, : [, : ACS]
전의 : [전 : N-의 : P][전의 : N]
커서 : [커서 : N][크 : V-어서 : F]
위치로 : [위치 : N-로 : P]
다시 : [다시 : ADV]
이동하시오 : [이동 : N-하 : SF-시 : PF-오 : F]
. : [. : ASC]

어휘중의성해소 결과(선택비율 0.001의 다중선택 방식 적용)

이 : [이 : DET]
줄 : [줄 : N][주 : V-ㄹ : F] : <- 미해결 어절
앞에 : [앞 : N-에 : P]
한 : [한 : DET]
행을 : [행 : N-을 : P]
추가하고 : [추가 : N-하 : SF-고 : F][추가 : N-하고 : P]
, : [, : ASC]
전의 : [전 : N-의 : P]
커서 : [커서 : N]
위치로 : [위치 : N-로 : P]
다시 : [다시 : ADV]
이동하시오 : [이동 : N-하 : SF-시 : PF-오 : F]
. : [. : ASC]

참 고 문 헌

1. K. Aijmer and B. Altenberg (eds), English Corpus Linguistics, Longman Inc., New York, 1991.
2. K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proc. of Second Conference on Applied Natural Language Processing*, pp. 136-143, Texas, USA, February 1988.
3. 이윤재, 최기선, 김길창, "한국어 문서 태깅 시스템," 정보과학회 봄 학술발표논문집, Vol. 20, No. 1, pp. 805-808, April 1993.
4. W. Gale, K. W. Church and D. Yarowsky, "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs," *Proc. of 30th Annual Meeting of the ACL*, pp. 249-256, Delaware, USA, June 1992.
5. 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 공학박사 학위논문, 1993.
6. S. J. DeRose, "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, Vol. 14, No. 1, pp. 31-39, December 1988.
7. C. G. de Marcken, "Prasing the LOB Corpus," *Proc. of 28th Annual Meeting of the ACL*, pp. 243-251, Pittsbugh, USA, June 1990.
8. K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29, March 1990.

李 夏 圭(Hagyu Lee) 정회원

1963年 10月 9日生

1987年 2月 : 서울대학교 컴퓨터공학과(공학사)
1989年 2月 : 서울대학교 대학원 컴퓨터공학과(공학석사)
1994年 2月 : 서울대학교 대학원 컴퓨터공학과(공학박사)

金 榮 澤(Yung Taek Kim) 정회원

1935年 10月 29日生

1963年 : 미국 Colorado대 전기과(공학석사)
1968年 : 미국 Uath대 전산과(공학박사)
1979年 : 미국 Purdue대와 Yale대 객원부교수
1980年 : 미국 Illinois대에서 컴파일러 연구
1981年 : 한국정보과학회 회장 역임
1990年 : 한국인지과학회 회장 역임
1970年 ~ 현재 : 서울대학교 컴퓨터공학과 교수