# 한국어 어휘 중의성 해소에서 어휘 확률에 대한 효과적인 평가 방법

이 하 규†


## 요 약

본 논문은 한국어 어휘 중의성 해소(lexical disambiguation)에서 어휘 확률(lexical probability) 평가 방법에 대해 기술하고 있다. 통계적 접근 방법의 어휘 중의성 해소에서는 일반적으로 말뭉치(corpus)로부터 추출된 통계 자료에 기초하여 어휘 확률과 문맥 확률(contextual probability)을 평가한다. 한국어는 어절별로 띄어쓰기가 이루어지므로 어절 단위로 어휘 확률을 적용하는 것이 바람직하다. 하지만 한국어는 어절의 다양성이 심하기 때문에 상당히 큰 말뭉치를 사용하더라도 어절 단위로는 어휘 확률을 직접 평가할 수 없는 경우가 다소 있다. 이러한 문제점을 극복하기 위해 본 연구에서는 어휘 분석 측면에서 어절의 유사성을 정의하고 이에 기반을 둔 한국어 어휘 확률 평가 방법을 제안한다. 이 방법에서는 어떤 어절에 대해 어휘 확률을 직접 평가할 수 없는 경우 이와 어휘 분석이 유사한 어절들을 통해 간접적으로 평가한다. 실험 결과 제안된 접근 방법이 한국어 어휘 중의성 해소에 효과적인 것으로 나타나고 있다.


# An Effective Estimation Method for Lexical Probabilities in Korean Lexical Disambiguation

Hagyu Lee†


## ABSTRACT

This paper describes an estimation method for lexical probabilities in Korean lexical disambiguation. In the stochastic approach to lexical disambiguation, lexical probabilities and contextual probabilities are generally estimated on the basis of statistical data extracted from corpora. It is desirable to apply lexical probabilities in terms of word phrases for Korean because sentences are spaced in the unit of word phrase. However, Korean word phrases are so multiform that there are more or less chances that lexical probabilities cannot be estimated directly in terms of word phrases though fairly large corpora are used. To overcome this problem, similarity for word phrases is defined from the lexical analysis point of view in this research and an estimation method for Korean lexical probabilities based on the similarity is proposed. In this method, when a lexical probability for a word phrase cannot be estimated directly, it is estimated indirectly through the word phrases similar to the given one. Experimental results show that the proposed approach is effective for Korean lexical disambiguation.


※이 논문은 1995년도 한림대학교 지원 학술연구조성비에 의하여 연구되었음.

† 정 회 원:한림대학교 컴퓨터공학과

논문접수:1996년 5월 1일, 심사완료:1996년 7월 4일

## 1. 서 론

통계적 접근 방법의 어휘 중의성 해소(lexical disambiguation) 혹은 품사 태깅(part-of-speech tagging)에서는 Shannon의 Noisy Channel 모형[6]에 근거를 두고 있는 경우가 많다[2]. 그리고 이러한 접근 방법에서 어휘 중의성 해소 모형의 두 가지 중요한 매개 변수인 어휘 확률(lexical probability)과 문맥 확률(contextual probability)은 일반적으로 말뭉치(corpus)에서 추출된 통계 자료에 기초하여 평가된다.

영어와 같이 단어별로 띄어쓰기를 하는 서구어의 어휘 중의성 해소에서는 대부분 단어 단위의 범주(품사)를 상정하고 있다. 그리고 어휘 확률로는 다음과 같이 단어 단위로 적용되는 두 가지 중 하나가 일반적으로 사용되고 있다[1, 2, 3, 4, 5]. 여기서 $w_i$는 $i$번째 단어를 나타내며, $wc_i$는 $w_i$의 단어 범주를 나타낸다.

$$\Pr(wc_i|w_i),\ \Pr(w_i|wc_i) \tag{1}$$

한국어는 어절별로 띄어쓰기가 이루어진다. 따라서 한국어의 어휘 중의성 해소에서는 어절 단위의 범주를 상정하는 접근 방식을 고려해 볼 수 있다[9, 10]. 그러면 다음과 같이 어휘 확률을 어절 단위로 적용 가능하다. 여기서 $wp_i$는 $i$번째 어절을 나타내며, $wpc_i$는 $wp_i$의 어절 범주를 나타낸다.

$$\Pr(wpc_i|wp_i),\ \Pr(wp_i|wpc_i) \tag{2}$$

그런데 이러한 접근 방식에서는 어절 범주 결정만으로는 한국어 어휘 중의성을 제대로 해소할 수 없는 경우가 종종 있다는 문제점이 있다. 예를 들어, 한국어에서 자주 사용되는 어절 중의 하나인 '나는'의 경우, 비록 'VE(동사 + 어말 어미)'와 같은 형태로 어절 범주가 결정되었다고 하더라도, '〈날/V, 는/E〉'인지 아니면 '〈나/V, 는/E〉'인지 구분할 수 없으므로 어휘 중의성이 완전히 해소되지 않는다.

한국어 어휘 중의성 해소를 위한 다른 접근 방식으로 형태소 단위의 범주를 상정하는 것을 고려해 볼 수 있다[8]. 그러면 다음과 같이 어휘 확률을 형태소 단위로 적용 가능하다. 여기서 $m_i$는 $i$번째 형태소를 나타내며, $mc_i$는 $m_i$의 형태소 범주를 나타낸다.

$$\Pr(mc_i|m_i),\ \Pr(m_i|mc_i) \tag{3}$$

이와 같은 접근 방식에서는 형태소 범주가 결정되면 어휘 중의성이 제대로 해소될 수 있다는 장점이 있지만 어휘 중의성 해소 과정에서 다음과 같은 어려운 점이 있다. 한국어에서는 동일한 어절에 대해 형태소의 개수가 다르게 어휘 분석이 되는 경우가 적지 않다. 형태소의 개수가 다르면 주어진 어절열 전체에 대해서 문맥 확률과 어휘 확률을 계산할 때, 보통 곱셈으로 적용되는 연산의 횟수가 형태소열에 따라 다를 수 있다. 따라서 이 방식에서는 개수가 다른 형태소열에 대해 비교의 공정성을 정확하게 유지하는 것이 어렵다. 그리고 형태소열에 대해 범주열을 결정하는 작업뿐만 아니라 주어진 어절열에 대해 형태소열을 결정하는 작업도 이루어져야 하므로 어휘 중의성 해소 과정이 복잡한 편이다. 실제로 한국어에 대해서 은닉 마르코프 모형(Hidden Markov Model)을 이용하여 이와 비슷한 접근 방식을 취한 경우가 있다[8]. 이 경우 형태소는 올바르게 분리되었지만 형태소 범주는 포함하지 않은 말뭉치를 사용하여 자율 학습 방법을 적용했는데, 어휘 중의성 해소의 정확성이 그렇게 높지 못한 것으로 나타나고 있다.

한국어 어휘 중의성 해소를 위한 또 다른 접근 방식으로 이상에서 언급된 문제점들을 극복하기 위해 제시된 TAIL-HEAD 공기(co-occurrence) 정보를 이용하는 방식이 있다[11, 12]. 개략적으로 말해 여기서 'HEAD'는 어절의 첫 단어를 지칭하며 'TAIL'은 어절의 마지막에 위치한 조사, 어미 등을 지칭한다. 그리고 'TAIL-HEAD 공기 정보'란 인접한 두 어절에 대해서 선행 어절의 TAIL과 후행 어절의 HEAD가 함께 나타나는 통계적 정보를 의미한다. 이 방식에서는 바로 이러한 TAIL-HEAD 공기 정보에 기반을 두고 이로부터 유도된 문맥 확률을 적용하고 있다. 그리고 어휘 확률은 어절 단위로 적용하되 어절 범주를 사용하는 대신 다음과 같이 해당 어절의 어휘 분석 결과로 주어지는 토큰열(token sequence)을 사용하고 있다. 여기서 $wp_i$는 $i$번째 어절을 나타내며, $ts_i$는 $wp_i$의 토큰열을 나타낸다.

$$\Pr(ts_i|wp_i) \tag{4}$$

이렇게 토큰열을 사용하는 경우 다음과 같은 형태의 어휘 확률을 적용하는 것은 적합하지 못하다.

$$\Pr(wp_i|ts_i) \quad (5)$$

왜냐하면, 어절 단위의 토큰열이 주어지면 해당 어절을 유일하게 식별할 수 있도록 토큰열을 표현할 때는 이 어휘 확률의 값이 항상 1이 되므로 어휘 중의성 해소에 전혀 도움이 되지 못하고, 그렇지 않을 때도 이 어휘 확률은 형태소의 축약이나 생략 현상이 일어난 경우와 이러한 현상이 일어나지 않은 경우 사이에서만 차이가 있으므로 어휘 중의성 해소에서 기여하는 정도가 별로 많지 않기 때문이다.

한국어 어휘 중의성 해소에서 TAIL-HEAD 공기 정보를 이용하는 위의 접근 방식은 다음과 같은 장점이 있다. 어휘 확률에서 토큰열을 사용하기 때문에 어절 단위의 범주를 상정하는 방식에 비해 어휘 중의성을 제대로 해소할 수 있다. 그리고 어절 단위의 어휘 확률과 TAIL-HEAD 공기 정보를 이용하는 문맥 확률을 적용하기 때문에 형태소 단위의 범주를 상정하는 방식에 비해 개수가 다른 형태소열에 대해서도 대체로 공정한 비교가 가능할 뿐만 아니라 어휘 중의성 해소 과정도 간단한 편이다. 또한 이러한 접근 방식은 실제 실험 결과에서도 앞에서 설명한 다른 방식에 비해 어휘 중의성 해소의 정확성이 비교적 높은 것으로 보고되고 있다[11, 12].

이상의 내용을 종합해 볼 때, 한국어 어휘 중의성 해소에서는 식 (4)와 같이 토큰열을 사용하여 어절 단위로 어휘 확률을 적용하는 것이 바람직한 것으로 보인다. 하지만 이와 같은 접근 방식에서는 어휘 확률을 직접 평가할 수 없는 경우가 다소 있다는 문제점이 있다. 한국어는 첨가어적 특성으로 인하여 영어등에 비하여 띄어쓰기 단위인 어절의 다양성이 매우 심한 편이다. 따라서 상당히 큰 말뭉치를 사용하더라도 어휘 확률을 어절 단위로 직접 평가하는 데 필요한 충분한 크기의 통계 자료가 추출되지 않을 가능성이 어느 정도 있다. 일반적으로 어휘 중의성 해소에서 어휘 확률은 문맥 확률에 비해 그 기여도가 크다[5, 11, 12]. 그러므로 통계 자료가 불충분하여 어휘 확률을 직접 평가할 수 없다고 해서 어휘 확률을 전혀 반영하지 않는다면 어휘 중의성 해소에서 높은 정확성을 기대하기 어렵다.[1)]

본 연구에서는 한국어 어휘 중의성 해소에서 어절 단위로 적용되는 어휘 확률이 지니고 있는 이상의 문제점을 극복하기 위해 어휘 분석적 측면에서 토큰열과 어절에 대해 유사성을 정의하고 이에 기반을 둔 어휘 확률 평가 방법을 제안한다. 한국어 말뭉치에 대해서 어휘 확률적 특성을 조사해 본 결과, 어절 및 토큰열에 대해 식 (4)의 어휘 확률을 직접 평가한 결과는 이와 어휘 분석이 유사한 어절 및 토큰열을 이용하여 간접적으로 평가한 결과와 매우 근사한 것으로 나타나고 있다. (3장 참고) 본 연구에서는 바로 이러한 한국어의 어휘 확률적 특성을 살려 주어진 어절 및 토큰열에 대해 식 (4)의 어휘 확률을 직접 평가할 수 없는 경우 이와 어휘 분석이 유사한 어절 및 토큰열들을 통하여 간접적으로 평가하는 접근 방식을 취하고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 토큰열과 어절의 어휘 분석 유사성에 대해 구체적으로 정의하며, 3장에서는 한국어의 어휘 확률적 특성에 대해 살펴보고 본 연구의 어휘 확률 평가 방법에 대해 설명한다. 그리고 4장에서 어휘 확률 평가 방법에 대한 실험 결과를 고찰하고, 5장에서 결론을 맺는다.

## 2. 어휘 분석 유사성

본 연구에서 도입하고 있는 어휘 분석 유사성이란 어절 단위의 두 토큰열 혹은 두 어절에 대해 어휘 분석적 측면에서 볼 때 유사성 여부를 판단하는 데 기준이 되는 개념이다. 어휘 분석 유사성에 대한 구체적인 정의에 앞서 먼저 본 논문에서 사용하는 몇 가지 기본적인 표기법을 살펴보면, $\langle tf_1/tc_1, tf_2/tc_2, \ldots, tf_n/tc_n \rangle$는 $n$개의 토큰으로 구성된 어절 단위의 토큰열을 나타낸다. 여기서 $tf_i$와 $tc_i$는 $i$번째 토큰의 어형과 범주를 각각 의미한다. 그리고 $lex(wp)$는 어절 $wp$의 어휘 분석 결과로 주어지는 토큰열의 집합을 나타

1) TAIL-HEAD 접근 방식에서 문맥 확률은 HEAD나 TAIL의 범주 정보를 이용하여 평가되기도 하므로 문맥 확률은 대부분의 경우 어느 정도 반영이 가능하다[11, 12].

낸다. 다음은 토큰열 집합의 표기 예를 보여준다. (이하 예에서 주어진 토큰의 범주에 대해서는 3장 참고)

lex(가옥이고)={〈가옥/N, 이/COP, 고/E〉,
〈가옥/N, 이고/P〉}
lex(건물이고)={〈건물/N, 이/COP, 고/E〉,
〈건물/N, 이고/P〉}
lex(가옥이며)={〈가옥/N, 이/COP, 며/E〉,
〈가옥/N, 이며/P〉}

본 연구에서는 두 가지 차원에서 어휘 분석 유사성을 정의하는데, 이를 위해 먼저 근사 토큰열 및 근사 토큰열 집합 그리고 토큰 범주열 및 토큰 범주열 집합을 다음과 같이 정의하고 있다.

정의 1 토큰열 $ts=\langle tf_1/tc_1, tf_2/tc_1,...,tf_n/tc_n\rangle$의 근사 토큰열 $approx(ts)$는 다음과 같이 정의한다. 여기서 λ는 토큰의 어형이 공백(null)임을 나타낸다.
$approx(ts)=\langle \lambda/tc_1, tf_2/tc_2,...,tf_n/tc_n\rangle$
그리고 어절 $wp$의 근사 토큰열 집합 $approx(wp)$는 다음과 같이 정의한다.
$approx(wp)=\{approx(ts)\mid ts\in lex(wp)\}$

정의 2 토큰열 $ts=\langle tf_1/tc_1, tf_2/tc_1,...,tf_n/tc_n\rangle$의 토큰 범주열 $cats(ts)$는 다음과 같이 정의한다.
$cats(ts)=\langle tc_1, tc_2,...,tc_n\rangle$
그리고 어절 $wp$의 토큰 범주열 집합 $cats(wp)$는 다음과 같이 정의한다.
$cats(wp)=\{cats(ts)\mid ts\in lex(wp)\}$

근사 토큰열의 경우 주어진 토큰열에서 첫번째 토큰의 어형만 공백으로 대치되고 나머지 부분은 주어진 토큰열과 동일하다. 그리고 토큰 범주열은 주어진 토큰열에서 토큰의 어형은 모두 제거하고 범주 정보만 추출한 것이다. 다음은 앞에서 예를 든 어절에 대해 근사 토큰열 집합과 토큰 범주열 집합을 보여준다.

approx(가옥이고)={〈λ/N, 이/COP, 고/E〉,
〈λ/N, 이고/P〉}
approx(건물이고)={〈λ/N, 이/COP, 고/E〉,
〈λ/N, 이고/P〉}
approx(가옥이며)={〈λ/N, 이/COP, 며/E〉,
〈λ/N, 이며/P〉}
cats(가옥이고)={〈N, COP, E〉, 〈N, P〉}
cats(건물이고)={〈N, COP, E〉, 〈N, P〉}
cats(가옥이며)={〈N, COP, E〉, 〈N, P〉}

다음은 토큰열과 어절에 대해 두 가지 차원의 어휘 분석 유사성, 즉 1차 어휘 분석 유사성과 2차 어휘 분석 유사성을 정의하고 있다.

정의 3 두 토큰열 $ts_i$와 $ts_j$의 근사 토큰열이 동일할 때, 즉 $approx(ts_i)=approx(ts_j)$일 때, $ts_i$와 $ts_j$ 사이에 1차 어휘 분석 유사성이 있다고 하며, 이 경우 다음과 같이 표기한다.
$ts_i\cong ts_j$
그리고 두 어절 $wp_i$와 $wp_j$의 근사 토큰열 집합이 동일할 때, 즉 $approx(wp_i)=approx(wp_j)$일 때, $wp_i$와 $wp_j$ 사이에 1차 어휘 분석 유사성이 있다고 하며, 이 경우 다음과 같이 표기한다.
$wp_i\cong wp_j$

정의 4 두 토큰열 $ts_i$와 $ts_j$의 토큰 범주열이 동일할 때, 즉 $cats(ts_i)=cats(ts_j)$일 때, $ts_i$와 $ts_j$ 사이에 2차 어휘 분석 유사성이 있다고 하며, 이 경우 다음과 같이 표기한다.
$ts_i\simeq ts_j$
그리고 두 어절 $wp_i$와 $wp_j$의 토큰 범주열 집합이 동일할 때, 즉 $cats(wp_i)=cats(wp_j)$일 때, $wp_i$와 $wp_j$ 사이에 2차 어휘 분석 유사성이 있다고 하며, 이 경우 다음과 같이 표기한다.
$wp_i\simeq wp_j$

이상의 정의에 따르면, 두 토큰열 혹은 두 어절 사이에 1차 어휘 분석 유사성이 있으면 항상 2차 어휘 분석 유사성이 있게 되지만 그 역은 성립하지 않는다. 그리고 일반적으로 1차 어휘 분석 유사성이 있는 토큰열이나 어절은 2차 어휘 분석 유사성만 있는 것에 비해 어휘 분석 측면에서 유사성 정도가 크다고 볼 수 있다. 다음은 앞에서 예를 든 어절에 대해 어휘 분

석 유사성 관계를 보여준다.

가옥이고 $\cong$ 건물이고 $\not\cong$ 가옥이며
가옥이고 $\simeq$ 건물이고 $\not\simeq$ 가옥이며

## 3. 어휘 확률적 특성과 어휘 확률 평가 방법

3장에서는 먼저 어휘 분석 유사성과 관련하여 한국어가 지니고 있는 어휘 확률적 특성을 고찰한 다음 본 연구에서 제안하고 있는 어휘 확률 평가 방법에 대해 기술하기로 한다.

한국어의 어휘 확률 특성 조사에 사용된 말뭉치는 약 25만 어절 규모로서 주로 전기, 전자, 전산 분야에 관련된 내용이다. 이 말뭉치는 연구 [7]에서 언급하고 있는 어휘 분석기의 처리 결과를 다음과 같은 본 연구의 토큰 범주 체계에 맞도록 변환한 후 어휘 중의성을 수동으로 해소하여 구축하였다. 이 말뭉치의 경우 어휘 분석 결과가 두 가지 이상인 중의적 어절은 전체 어절의 34.6%였으며 중의적 어절의 평균 토큰열 수는 2.6개로 나타났다.

| | |
|---|---|
| PRON: 대명사 | P: 조사 - 서술격 조사 제외 |
| N: 명사 | E: 어말 어미 - 명사형 어미 제외 |
| NUM: 수사 | PFE: 선어말 어미 |
| V: 동사 | COP: 서술격 조사 |
| ADJ: 형용사 | NMZ: 명사형 어미 |
| ADV: 부사 | SFX: 접미사 |
| DET: 관형사 | AUX-V: 보조 동사 - 붙여쓴 경우 |
| EXCL: 감탄사 | AUX-A: 보조 형용사 - 붙여쓴 경우 |
| PUNC: 문장 부호 | |

어휘 확률 특성 조사에서는 앞에서 언급한 말뭉치에 나타난 중의적 어절의 각 토큰열-어절 쌍에 대해서 역시 이 말뭉치에서 추출된 통계 정보를 이용하여 다음 식 (6)의 세 가지 어휘 확률 평가식을 각각 적용하였을 때 실제로 계산된 평가값들 사이의 차이를 관측하였다. 여기서 *freq*는 말뭉치에서 나타난 빈도를 의미하고 *ts*와 *wp*는 각각 토큰열과 어절을 나타낸다. 기본 어휘 확률 평가식에서는 식 (4)의 어휘 확률을 직접적으로 평가하며, 1차 및 2차 근사 어휘 확률 평가식에서는 2장에서 정의된 1차 및 2차 어휘 분석 유사성 관계를 이용하여 간접적으로 평가하고 있다.

기본 어휘 확률 평가식:

$$Pr_b(ts|wp)=\frac{freq(ts,\ wp)}{freq(wp)}$$

1차 근사 어휘 확률 평가식:

$$Pr_1(ts|wp)=\frac{freq(ts',\ wp')}{freq(wp')} \qquad (6)$$

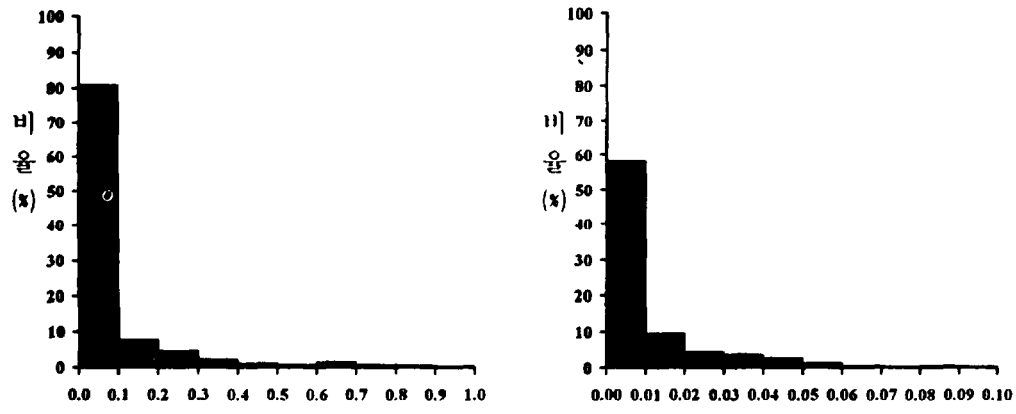
2차 근사 어휘 확률 평가식:

$$Pr_2(ts|wp)=\frac{freq(ts'',\ wp'')}{freq(wp'')}$$

(여기서, $wp' \cong wp$, $ts' \cong ts$, $wp'' \simeq wp$, $ts'' \simeq ts$)

(그림 1)과 (그림 2)는 각각 기본 어휘 확률 평가식과 1차 근사 어휘 확률 평가식 간의 차이 및 기본 어휘 확률 평가식과 2차 근사 어휘 확률 평가식 간의 차이를 구간별 백분 비율 분포로 보여준다. 이 두 그림에서 (a)는 전체 구간에 대한 분포이며, (b)는 차이가 0 이상 0.1 이하인 구간에 대한 분포이다. 기본 어휘 확률 평가식과 1차 근사 어휘 확률 평가식의 차이는 전체 토큰열-어절 쌍 가운데서 81.8%가 0.1 이하이며, 58.1%가 0.01 이하인 것으로 나타났다. 그리고 기본 어휘 확률 평가식과 2차 근사 어휘 확률 평가식의 차이는 0.1 이하인 토큰열-어절 쌍이 전체의 58.6%이며, 0.01 이하인 것이 34.9%로 나타났다.

이상의 어휘 확률 특성 조사 결과로부터 1차 근사 어휘 확률 평가식은 기본 어휘 확률 평가식과 매우 근사하며, 2차 근사 어휘 확률 평가식도 대체로 근사하다는 것을 알 수 있다. 이는 말뭉치로부터 추출된 통계 정보가 부족하여 어휘 확률을 직접 평가하기 어려운 경우, 어휘 분석 유사성이 있는 어절 및 토큰열을 통해 간접적으로 평가하여도 비교적 정확한 어휘 확률 값을 구할 수 있다는 것을 의미한다. 본 연구에서는 바로 이러한 한국어의 어휘 확률적 특성을 고려하여, 식 (4)의 어휘 확률에 대해 다음 식 (7)과 같은 평가 방법을 제안한다.

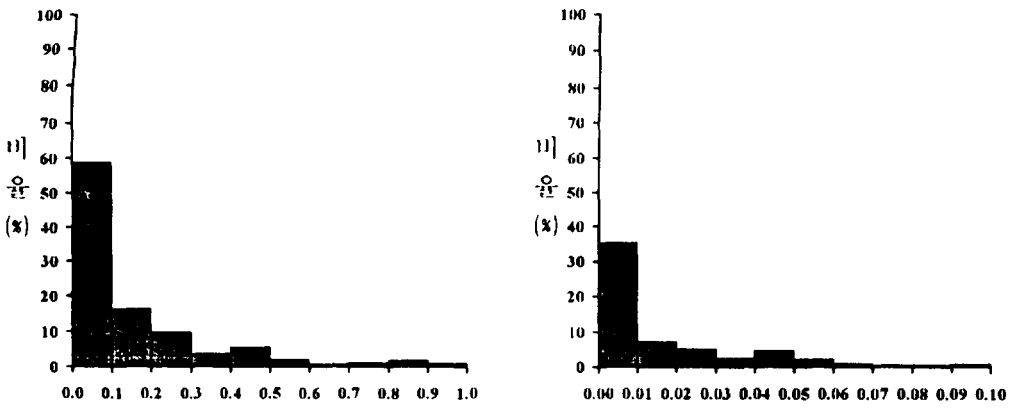



$d_{b1} = |Pr_b(ts|wp) - Pr_1(ts|wp)|$

(a) $0 \le d_{b1} \le 1$ (b) $0 \le d_{b1} \le 0.1$

(그림 1) 기본 어휘 확률 평가식과 1차 근사 어휘 확률 평가식의 차이

(Fig. 1) Differences between the Basic Lexical Probability Estimation Formula and the 1st Order Approximate Lexical Probability Estimation Formula




$d_{b2} = |Pr_b(ts|wp) - Pr_2(ts|wp)|$

(a) $0 \le d_{b2} < 1$ (b) $0 \le d_{b2} \le 0.1$

(그림 2) 기본 어휘 확률 평가식과 2차 근사 어휘 확률 평가식의 차이

(Fig. 2) Differences between the Basic Lexical Probability Estimation Formula and the 2nd Order Approximate Lexical Probability Estimation Formula

$$\Pr(ts|wp) \approx \begin{cases} Pr_b(ts|wp) & \text{if} \quad freq(wp) \ge \alpha \\ Pr_1(ts|wp) & \text{else if } freq(wp') \ge \beta \\ Pr_2(ts|wp) & \text{else if } freq(wp'') \ge \gamma \\ \delta & \text{otherwise} \end{cases} \quad (7)$$

(여기서, $wp' \cong wp$, $wp'' \simeq wp$)

여기서 $\delta$는 어휘 확률을 평가할 수 없을 때 기본적으로 주어지는 어휘 확률 값을 나타낸다. 그리고 $\alpha$, $\beta$, $\gamma$는 각각 $wp$, $wp'$, $wp''$에 대한 임계 빈도를 나타낸다. 말뭉치로부터 추출된 $wp$, $wp'$ 혹은 $wp''$의 빈도가 너무 적은 경우에는 해당 어휘 확률 평가식의 계산 결

과를 신뢰할 수 없으므로 어휘 확률을 배제하고 문맥 확률만 이용하여 어휘 중의성을 해소하는 것이 정확성이 높을 가능성이 많다. 따라서 $wp$, $wp'$, $wp''$ 각각에 대해 어휘 중의성 해소의 정확성을 최대화시키는 임계 빈도를 실험을 통해 구하고 $wp$, $wp'$, $wp''$의 빈도가 해당 임계 빈도 이상인 경우에만 관련된 어휘 확률 평가식을 적용하는 것이 타당성이 있다. 그리고 한국어의 어휘 확률 특성 조사에서 1차 근사 어휘 확률 평가식이 2차 근사 어휘 확률 평가식에 비해 기본 어휘 확률 평가식에 보다 가까운 것으로 나타나고 있다. 따라서 기본 어휘 확률 평가식, 1차 근사 어휘 확률 평가식, 2차 근사 어휘 확률 평가식의 순서로 어휘 확률 평가식을 단계적으로 적용하는 것이 적합하다.

본 연구에서는 이상과 같은 점을 고려하여 주어진 어절의 빈도 $freq(wp)$가 임계 빈도 $\alpha$ 이상이면 기본 어휘 확률 평가식 $Pr_0(ts|wp)$을 적용하고, 그렇지 않고 주어진 어절과 1차 어휘 분석 유사성이 있는 어절의 빈도 $freq(wp')$가 임계 빈도 $\beta$ 이상이면 1차 근사 어휘 확률 평가식 $Pr_1(ts|wp)$을 적용하고, 그렇지 않고 2차 어휘 분석 유사성이 있는 어절의 빈도 $freq(wp'')$가 임계 빈도 $\gamma$ 이상이면 2차 근사 어휘 확률 평가식 $Pr_2(ts|wp)$을 적용하며, 그렇지 않은 경우에는 기본적으로 주어지는 어휘 확률 값 $\delta$를 적용하여 사실상 어휘 확률을 반영하지 않는 접근 방법을 취하고 있다. 한국어에서 어절 단위의 어휘 확률을 직접 평가하기 어려운 경우, 이와 같이 어휘 분석이 유사한 어절과 토큰열을 통해 간접적으로 평가하면, 엄청나게 큰 말뭉치를 사용하지 않더라도 거의 대부분 비교적 정확하게 어휘 확률이 반영되므로 어휘 중의성 해소의 정확성을 향상시킬 수 있다.

## 4. 실 험

4장은 본 연구에서 제안하고 있는 한국어 어휘 확률 평가 방법의 타당성을 검증하고 실제 어휘 중의성 해소에서 어느 정도 효과가 있는지 알아보기 위해 수행한 실험 및 결과에 대해 설명한다. 먼저 실험 환경을 살펴보면 다음과 같다. 어휘 분석기와 토큰의 범주 체계 및 통계 정보 추출에 사용된 말뭉치는 3장에서 언급한 어휘 확률 특성 조사에서 사용된 것과 동일하다. 그리고 이 말뭉치와는 별도로 마련된 약 3만 어절 규모의 테스트 자료에 대해서 어휘 중의성 해소의 정확성을 측정하였는데, 통계 정보 추출에 사용된 말뭉치와 마찬가지로 테스트 자료의 내용도 주로 전기, 전자, 전산 분야에 관련된 것이다. 테스트 자료의 경우 어휘 분석 결과가 두 가지 이상인 중의적 어절은 전체 어절의 35.0%였으며, 중의적 어절의 평균 토큰열 수는 2.6개로 나타났다. 이 실험에서 사용된 임계 빈도 $\alpha$, $\beta$, $\gamma$는 각각 3, 5, 6으로서, 이들 임계 빈도는 연구 [11]의 한국어 어휘 중의성 해소 모형에 식 (7)의 어휘 확률 평가 방법을 적용하였을 때 어휘 중의성 해소의 정확성이 가장 높게 나타난 경우의 값이다. 그리고 기본적으로 주어지는 어휘 확률 값 $\delta$는 1을 사용하였다.[2)]

〈표 1〉은 식 (4)의 어휘 확률에 대해 다음 식 (8)에서 주어진 4 가지 어휘 확률 평가 방법들 간에 비교 실험 결과를 보여준다. 여기서 방법 A는 연구 [11, 12]에서 사용된 기존의 어휘 확률 평가 방법으로서 식 (6)에서 주어진 기본 어휘 확률 평가식이 사용되고 있다. 방법 B와 C에서는 각각 1차 및 2차 근사 어휘 확률 평가식을 사용하여 어휘 확률을 평가한다. 그리고 방법 D는 식 (7)에서 주어진 어휘 확률 평가 방법으로서 본 연구에서 최종적으로 제안하고 있는 방법이다. 이 방법에서는 기본 어휘 확률 평가식과 1차 및 2차 근사 어휘 확률 평가식이 모두 사용되고 있다.

$$
\begin{aligned}
&\text{방법 A:}\\
&\Pr(ts|wp) \approx \begin{cases} Pr_0(ts|wp) & \text{if } freq(wp) \geq \alpha \\ \delta & \text{otherwise} \end{cases}\\
&\text{방법 B:}\\
&\Pr(ts|wp) \approx \begin{cases} Pr_1(ts|wp) & \text{if } freq(wp') \geq \beta \\ \delta & \text{otherwise} \end{cases}\\
&\text{방법 C:}\\
&\Pr(ts|wp) \approx \begin{cases} Pr_2(ts|wp) & \text{if } freq(wp'') \geq \gamma \\ \delta & \text{otherwise} \end{cases}\\
&\text{방법 D:}\\
&\text{식 (7)에서 주어진 어휘 확률 평가 방법}\\
&\text{(여기서, } wp' \cong wp,\ wp'' \simeq wp\text{)}
\end{aligned}
\tag{8}
$$

〈표 1〉에서 두번째 열은 테스트 자료의 중의적 어

2) $\delta$ 값이 어휘 중의성 해소 결과에 실제로 영향을 주지는 않으므로 임의의 양수 값을 취해도 결과는 동일하게 된다.

〈표 1〉 어휘 확률 평가 방법의 비교

〈Table 1〉 Comparison of the Lexical Probability Estimation Methods

| 어휘 확률 평가 방법 | 적용된 중의적 어절의 비율 | 어휘 확률의 중의성 해소 정확성 | |
|---|---|---|---|
| | | 적용된 중의적 어절 | 전체 중의적 어절 |
| 방법 A | 84.1% | 95.3% | 86.3% |
| 방법 B | 97.5% | 93.2% | 91.8% |
| 방법 C | 99.8% | 87.5% | 87.4% |
| 방법 D | 99.8% | 94.8% | 94.7% |

(말뭉치 크기:25만 어절)

〈표 2〉 말뭉치 크기별 비교

〈Table 2〉 Comparison by the Corpus Size

| 말뭉치 크기 | 5만 어절 | 10만 어절 | 15만 어절 | 20만 어절 | 25만 어절 |
|---|---|---|---|---|---|
| (A) | 74.4% | 78.7% | 82.0% | 84.5% | 86.3% |
| (D) | 88.3% | 90.5% | 92.2% | 93.6% | 94.7% |
| (D)-(A) | 13.9% | 11.8% | 10.2% | 9.1% | 8.4% |

(A): 전체 중의적 어절에 대한 방법 A의 어휘 중의성 해소 정확성

(D): 전체 중의적 어절에 대한 방법 D의 어휘 중의성 해소 정확성

절 가운데 말뭉치에서 추출된 빈도 $freq(wp)$, $freq(wp')$, $freq(wp'')$가 해당 임계 빈도 이상으로 나타나 어휘 확률 평가식 $Pr_b(ts|wp)$, $Pr_1(ts|wp)$, $Pr_2(ts|wp)$가 적용된 중의적 어절의 비율이다. 세번째 열은 이와 같이 어휘 확률 평가식이 적용된 어절에 대한 평균적인 어휘 중의성 해소의 정확성이다. 그리고 네번째 열은 어휘 확률 평가식이 적용되지 않은 어절의 경우 무작위로 하나의 토큰열을 선택하였을 때 중의적 어절 전체에 대한 정확성을 보여준다.

이상의 실험 결과를 살펴보면, 기존의 방법 A는 적용된 어절에 대해서는 다른 방법에 비해 정확성이 우수하지만 적용된 어절의 비율이 높지 않기 때문에 전체적인 정확성이 가장 낮은 것으로 나타났다. 1차 근사 어휘 확률 평가식을 사용한 방법 B는 적용된 어절의 비율도 비교적 높고 전체적인 정확성도 상당히 높은 것으로 나타났다. 2차 근사 어휘 확률 평가식을 사용한 방법 C는 대부분의 어절에 대해 적용되었지만 적용된 경우의 정확성이 다른 방법에 비해 낮기 때문에 전체적인 정확성이 떨어지는 것으로 나타났다. 이에 반해 본 연구에서 제안하고 있는 방법 D는 방법 C와 마찬가지로 대부분의 어절에 대해 어휘 확률 평가식이 적용되었고 적용된 경우의 정확성도 상당히 높기 때문에 전체적인 어휘 중의성 해소의 정확성이 가장 우수한 것으로 나타났다. 이와 같은 실험 결과로부터 본 연구의 접근 방법이 어절 단위로 적용되는 식 (4)의 한국어 어휘 확률 평가 방법으로 보다 적합하다는 것을 알 수 있다.

〈표 2〉는 기존의 방법 A와 본 연구의 방법 D 사이에 비교 실험 결과를 말뭉치의 크기별로 보여준다. 말뭉치가 작은 경우, 방법 A는 기본 어휘 확률 평가식만 사용하므로 정확성이 많이 감소하지만 방법 D는 1차 및 2차 근사 어휘 확률 평가식도 사용하므로 정확성이 비교적 적게 감소하는 것으로 나타났다. 그리고 말뭉치가 커질수록 기본 어휘 확률 평가식이 적용되는 어절의 비율이 증가하기 때문에 방법 A와 방법 D의 정확성 차이가 점차 줄어드는 것으로 나타났다. 이로부터 본 연구의 접근 방법은 말뭉치가 작은 경우에 특히 효과적이고, 말뭉치의 규모가 증가하면 다소 효과가 감소함을 알 수 있다.

〈표 3〉은 연구 [11]에서 제시된 TAIL-HEAD 접근 방식의 한국어 어휘 중의성 해소 모형에 대해 기존의 어휘 확률 평가 방법 A와 본 연구에서 제안된 어휘

확률 평가 방법 D를 각각 적용하였을 경우에, 테스트 자료의 중의적 어절과 전체 어절에 대한 어휘 중의성 해소의 정확성을 보여준다. 실험 결과 중의적 어절에 대한 방법 A와 방법 D의 어휘 중의성 해소 오류가 각각 4.9%와 2.3%인 것으로 나타나, 본 연구의 방법을 적용하였을 때 53.1%의 오류 감소율을 보였다. 이와 같은 실험 결과로부터 본 연구의 접근 방법이 문맥 확률까지 고려된 실제 한국어 어휘 중의성 해소에서도 정확성 향상 효과가 크다는 것을 알 수 있다.

〈표 3〉 TAIL-HEAD 접근 방식에서의 비교
〈Table 3〉 Comparison in the TAIL-HEAD Approach

| 어휘 확률 평가 방법 | 어휘 중의성 해소의 정확성 | |
|---|---|---|
| | 중의적 어절 | 전체 어절 |
| 방법 A | 95.1% | 98.3% |
| 방법 D | 97.7% | 99.2% |

(말뭉치 크기:25만 어절)

## 5. 결 론

한국어는 어절별로 띄어쓰기가 이루어지므로 어휘 중의성 해소에서 어휘 확률을 어절 단위로 적용하는 것이 바람직하다. 하지만 한국어는 어절의 다양성이 심하기 때문에 상당히 큰 말뭉치를 사용하더라도 어휘 확률을 어절 단위로 직접 평가하기 어려운 경우가 있다. 이러한 문제점을 극복하기 위해 본 연구에서는 토큰열 및 어절에 대해서 두 가지 어휘 분석 유사성을 정의하고, 어휘 분석 유사성과 관련하여 한국어가 지니고 있는 어휘 확률적 특성을 고려하여, 주어진 어절과 토큰열에 대해 어휘 확률을 직접 평가할 수 없는 경우 이와 어휘 분석이 유사한 어절 및 토큰열을 통해 어절 단위의 어휘 확률을 간접적으로 평가하는 방식을 제안하고 있다. 실험 결과 본 연구에서 제안된 어휘 확률 평가 방법이 한국어 어휘 중의성 해소의 정확성을 향상시키는 데 많은 효과가 있으며, 말뭉치의 규모가 작은 경우에 특히 효과적인 것으로 나타났다.

본 연구의 접근 방법은 어절 단위의 범주를 상정하는 한국어 어휘 중의성 해소 방식[9, 10]에서 사용되는 식 (2)와 같은 형태의 어휘 확률을 평가할 때에도 비슷하게 응용 가능한 것으로 보인다. 그리고 말뭉치가 커질수록 본 연구의 어휘 확률 평가 방법은 다소 효과가 감소하는 것으로 나타나고 있다. 따라서 앞으로 식 (2)와 같은 형태의 어휘 확률 평가 방법에 관한 연구와 말뭉치가 보다 큰 경우에 본 연구의 접근 방법이 어느 정도 효과가 있을 것인지에 대한 검토가 있어야 할 것으로 사료된다.

## 참 고 문 헌


[1] Church, K. W., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proc. of Second Conference on Applied Natural Language Processing*, pp. 136-143, 1988.

[2] Church, K. W. and Mercer, R. L., "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, Vol. 19, No. 1, pp. 1-24, 1993.

[3] De Marcken, C. G., "Parsing the LOB Corpus," *Proc. of 28th Annual Meeting of the ACL*, pp. 243-251, 1990.

[4] DeRose, S. J., "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, Vol. 14, No. 1, pp. 31-39, 1988.

[5] Franz. A., "An Exploration of Stochastic Part-of-Speech Tagging," *Proc. of NLPRS '95*, Vol. 1, pp. 217-222, 1995.

[6] Shannon, C. E., "The Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, pp. 379-656, 1948.

[7] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 공학박사 학위 논문, 1993.

[8] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅," 정보과학회논문지, 제22권, 제1호, pp. 136-146, 1995.

[9] 박혜준, 윤준태, 송만석, "말뭉치 품사꼬리달기 시스템 구현," 정보과학회 봄 학술발표논문집, pp. 829-832, 1994.

[10] 이운재, 최기선, 김길창, "한국어 문서 태깅 시스템," 정보과학회 봄 학술발표논문집, pp. 805-808,

1993.

[11] 이하규, "Tail-Head 공기 정보를 이용한 한국어 어휘 중의성 해소," 자연어처리 기술 보고서 NLP-TECH-96-1, 한림대학교 컴퓨터공학과 자연어처리 연구실, pp. 1-13, 1996.

[12] 이하규, 김영택, "통계정보에 기반을 둔 한국어 어휘중의성해소," 한국통신학회지, 제19권, 제2호, pp. 265-275, 1994.

**이 하 규**

1987년 서울대학교 전자계산기공학과(학사)

1989년 서울대학교 대학원 전자계산기공학과(석사)

1994년 서울대학교 대학원 컴퓨터공학과(박사)

1995년~현재 한림대학교 컴퓨터공학과 전임강사

관심분야: 한국어정보처리, 자연어처리, 정보검색