KIPS Tr. Software and Data Eng.
Vol.2, No.12 pp.881~888 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.12.881

Construction of an Efficient Pre-analyzed Dictionary for Korean Morphological Analysis

Sujeong Kwak† · Bogyum Kim†† · Jae Sung Lee†††

ABSTRACT

A pre-analyzed dictionary is used to increase the speed and the accuracy of morphological analyzers and to decrease the over-generation. However, if the dictionary includes 'Insufficiently-analyzed word-phrases', which do not include all the possible analysis of the word-phrase, it may cause the decrease of the analysis accuracy. In this paper, we measure the accuracy changes according to the number of word-phrase frequency and the size changes of corpus by Sejong corpus. And performance of integrate system(SMA with pre-dictionary) is highest when sufficient analysis rate of pre-dictionary is more than 99.82%. Also pre-dictionary is constructed with word-phrase that frequency more than 32(64) when size of corpus is 1,600,000(6,300,000) word-phrase.

Keywords : Morphological Analysis, Pre-compiled Dictionary, Insufficiently-analyzed word-phrase, Over-generation

한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법

곽 수 정† · 김 보 겸†† · 이 재 성†††

요 약

기분석 사전은 형태소 분석기의 속도와 정확도를 향상시키고, 과분석을 줄이기 위해 사용된다. 하지만 기분석 사전에 저장된 어절 중에 저장된 형태소 분석 결과가 부족한 어절, 즉 불충분 분석 어절이 존재할 경우 오히려 형태소 분석기의 정확도를 떨어뜨리는 원인으로 작용할 수 있다. 본 논문에서는 세종 형태 분석 말뭉치(문어체, 2011)를 이용해 말뭉치의 크기와 어절 빈도의 변화에 따라 사전의 정답 제시율이 변화하는 양상을 측정하였다. 그리고 통계기반의 형태소 분석기인 SMA와 기분석 사전을 결합한 통합 시스템을 구성하여 기분석 사전의 충분 분석률이 99.82% 이상일 때 시스템 전체 성능이 향상되는 것을 확인하였다. 또한 160만 어절의 말뭉치를 이용할 때는 32회 이상 출현한 어절로, 630만 어절로 구성된 말뭉치를 이용할 때는 64회 이상 출현한 어절로 사전을 구성하는 것이 통합 시스템의 성능을 가장 높게 할 수 있었다.

키워드 : 형태소분석, 기분석 사전, 불충분분석어절, 과분석

1. 서 론

형태소 분석이란 자연언어처리의 가장 기본적인 단계로, 어절을 의미를 가지는 가장 작은 단위인 형태소로 분리하고, 품사를 찾아내는 것이다[1]. 한국어는 교착어의 특성을 가지며, 다양한 음운 현상이 발생하기 때문에 형태소 분석 과정이 매우 복잡하다. 따라서 빠른 속도의 형태소 분석을 위해 기분석 사전을 이용한 형태소 분석이 제안되어 왔다[2, 3].

기분석 사전은 어절에 대한 형태소 분석을 미리 수행하여 저장해 둔 것으로 이를 사용하면 실행 시간에 특별한 연산 과정 없이 사전 탐색만으로 형태소 분석이 가능하기 때문에 처리가 간단하고 빠르다. 하지만 한국어 어절의 종류는 사실상 무한하기 때문에 모든 어절을 기분석 사전에 넣을 수 없다[3]. 따라서 일반적으로 기분석 사전은 규칙이나 통계에 기반한 형태소 분석기와 결합하여 사용된다. 이 경우, 우선 입력된 어절을 기분석 사전에서 탐색하고, 그 어절이 있을 경우 미리 분석된 결과를 그대로 출력한다. 기분석 사전에 입력 어절이 없을 경우에만 형태소 분석기를 이용해 분석을 수행한다[4, 5].

기분석 사전은 일반적으로 대용량의 품사 부착 말뭉치를 이용해 구축한다. 이것은 이미 정답으로 확인된 분석 결과만 출력하기 때문에 과분석을 줄일 수 있다는 장점이 있다[3]. 하지만 이와 모순되게 중의성을 가지는 어절의 분석 결

※ 이 논문은 2012년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.
† 준 회 원 : 충북대학교 정보산업공학협동과정 석사과정
†† 준 회 원 : 충북대학교 디지털정보융합학과 박사과정
††† 종신회원 : 충북대학교 디지털정보융합학과 교수
논문접수 : 2013년 8월 9일
수 정 일 : 1차 2013년 9월 9일, 2차 2013년 10월 22일
심사완료 : 2013년 10월 22일
* Corresponding Author : Jae Sung Lee(jasonlee@cbnu.ac.kr)

과 중 일부만 말뭉치에 출현한 경우 원하는 분석 결과가 포함되지 않아 형태소 분석기 전체의 정답 제시율을 떨어뜨릴 수 있다. 따라서 기분석 사전을 사용하기에 앞서 분석 결과가 부족한 어절을 제거하거나, 이러한 어절에 의한 사전의 오류율을 파악할 필요가 있다.

본 논문에서는 말뭉치의 크기와 말뭉치를 필터링할 어절의 출현 빈도 기준을 다르게 하여 사전을 구성하였다. 그리고 사전의 성능을 측정하여 이것이 형태소 분석기에 미칠 영향에 대해서 추정하고, 사전 구성 시 필요한 최소한의 말뭉치 크기를 제안한다.

2장에서는 기분석 사전에 대한 관련 연구를 살펴보고, 3장에서는 분석 결과가 부족한 어절을 사전에서 제거하는 방안에 대하여 논의한다. 4장에서는 어절 출현 빈도와 기분석 사전 크기에 따른 성능을 측정하고, 5장에서는 기분석 사전과 통계기반 형태소 분석기를 결합한 통합 시스템의 성능을 분석한다. 이어 6장에서 결론을 맺는다.

2. 관련 연구

기분석 사전은 주로 대용량의 품사 부착 말뭉치의 고빈도 어절로 구축되어 형태소 분석기의 속도를 향상시키기 위해 사용되어왔다. [2]는 350만 어절의 원시 말뭉치에서 평균 적중률이 75.1%에 이르는 15만 어절을 추출하였고, [6]은 다양한 기관에서 구성한 대량의 말뭉치에서 공동으로 출현한 약 18만의 고빈도 어절을 추출해, 적중률이 81.5%임을 보였다. 이를 통해 고빈도 어절을 저장한 기분석 사전을 형태소 분석에 활용하는 것이 효율적임을 증명하였다.

기분석 사전은 다양한 방법으로 형태소 분석기에 활용되었다. 이것은 크게 전체 어절을 저장하여 사용하는 어절 단위의 기분석 사전과 어절의 일부분에 대한 분석만을 저장해 사용하는 부분 어절 단위 기분석 사전으로 나눌 수 있다[7].

어절 단위 기분석 사전은 이제까지 살펴본 것과 같이 사전 탐색만으로 형태소 분석이 가능하기 때문에 속도가 빠르며 올바른 분석 결과만 저장하여 과분석이 없다는 장점이 있다. 하지만 모든 어절에 대한 분석 결과를 저장할 수 없기 때문에 어절 단위 기분석 사전만 사용해 형태소 분석을 하는 것은 한계가 있다[3].

부분 어절 단위 기분석 사전은 문법 형태소 사전, 선어말어미 사전 등의 특정 부분만 저장해 사용하는 것[8]과 어절의 모든 부분을 조각으로 잘라 인접 정보와 부분 어절의 기분석 정보만으로 형태소 분석이 가능하게 한 것[5, 7], 실질 형태소를 개방어로 두고 문법 형태소로 구성된 어절 패턴 사전을 구축해 형태소 분석기로 활용한 것[9] 등이 있다. 이러한 부분 어절 단위 기분석 사전은 모든 어절을 분석할 수 없는 기분석 사전의 한계를 극복하였다. 하지만 가능한 결합이나 적용 패턴이 너무 많아 과분석이 발생하는 문제가 있다.

본 논문에서는 어절 단위 기분석 사전에 초첨을 맞춰 사전을 구성하는 말뭉치의 크기와 필터링 기준에 따른 기분석 사전의 성능을 평가하고, 형태소 분석기과 통합하여 적용해 보았다.

3. 불충분 분석 어절 제거에 대한 논의

한국어 분석에 있어서 어절의 중의성 해결은 매우 중요한 문제 중 하나이다[10-12]. 어절의 중의성은 문맥을 파악할 수 있는 문장 단위 이상에서 해결 가능한 문제로 어절 단위의 분석을 시행하는 형태소 분석기에서 이것을 해결하는 것은 매우 어렵다. 따라서 형태소 분석기는 중의성을 가지는 모든 분석 결과를 출력해 상위 단계에서 문맥에 맞는 것을 선정할 수 있도록 해야 한다[12]. 이는 형태소 분석기에서 사용하는 기분석 사전에도 적용되어, 가능한 모든 결과가 출력되어야 한다. 그런데 사전에 저장된 중의성 어절 중 분석의 일부만 저장된 어절이 존재한다면 이는 정확도를 떨어뜨리는 요인이 될 수 있다. 예를 들어, '나는'에 대한 형태소 분석은 '나/NP+는/JKC', '날/VV+는/EC', '나/VV+는/EC'가 모두라고 가정하자. 만약 기분석 사전에 위의 3가지 분석 중 일부만 포함되어 '나/VV+는/EC'가 없다고 가정하면, 이 분석은 항상 실패하게 된다. 이런 경우의 어절, 즉, 모든 경우의 분석을 포함하지 않은 어절을 불충분 분석 어절이라고 하자. 본 논문에서는 기분석 사전에서 나타난 불충분 분석 어절의 특성을 파악하여, 사전에 저장된 불충분분석 어절의 수를 조절하고, 기존의 형태소 분석기와 통합하여 효율적으로 사용할 수 있는 방법에 대해서 논의한다.

불충분 분석 어절의 특징을 확인하기 위해 먼저 말뭉치를 이용해 사전을 구축하고, 불충분 분석 어절을 파악할 필요가 있다. 이에 세종 형태 분석 말뭉치(2011, 문어체)[13]에서 오류와 특수기호를 제거한 700만여 어절을 실험에 사용하였다. 말뭉치의 90%는 사전으로 10%는 평가를 위한 데이터로 하여 평가 데이터의 어절 중 사전에서 탐색되지만 올바른 형태소 분석에 실패한 어절을 불충분 분석으로 보고, 특징 파악에 사용하였다. 평가 데이터의 192,997 어절 중 139,677어절이 적중하였으며, 이중 3,169어절이 불충분 분석 어절 이었다.

3.1 동형이의 문법 형태소를 이용한 어절 필터링

3,169개의 불충분 분석 어절 중 151(4.76%)개의 어절은 원형 복원이나 분리 결과가 정답과 다른 형태 중의성에 의한 불충분 분석 어절이고, 2,972(93.78%)개의 어절은 원형 복원과 분리 결과는 같지만 부착된 품사가 정답과 다른 품사 중의성에 의한 불충분 분석 어절이다. 그리고 46(1.46%)개의 어절은 이 두 가지 경우를 모두 포함하였다. 품사 중의성에 의한 불충분 분석 어절이 93.78%에 이르므로 이것을 사전에서 제거한다면 간단하게 기분석 사전의 성능을 향상 시킬 수 있을 것이다. 이에 품사 중의성에 의한 불충분 분석 어절 100개를 임의로 뽑아 그 특성을 직접 확인하였다(Table 1).

Table1. Types of POS Ambiguity

type	consistency error	homograph		
		full- full	full - empty	empty- empty
the number of word-phrase(not considering the frequency)	9	11	3	78

추출된 100개의 어절 중 78개의 어절이 두개의 문법 형태소가 동형이의 관계에 있어 불충분 분석 어절이 되었다. Fig. 1은 그 예이다.

활용이 **가능해** 개발 경쟁이 치열하다.
가능/NNG+하/XSA+**아/EC**

자네라면 **가능해.**
가능/NNG+하/XSA+**아/EF**

Fig. 1. Example of homograph empty morpheme

예제를 살펴보면 문장에서 어절이 나타난 위치에 따라서 문법 형태소의 품사가 달라지는데 문법 형태소는 그 수가 한정되어 있고 어느 정도의 규칙을 가지고 위치에 따라 다른 의미를 가지기 때문에 불충분 분석 어절을 제거하는 요소로 이용할 수 있다.

가장 많이 혼동 되는 문법 형태소의 품사는 예제에서 나타난 연결 어미(EC)-종결 어미(EF) 외에도 부사격 조사(JKB)-접속 조사(JC), 주격 조사(JKS)-보격 조사(JKC)가 있다. 이렇게 혼동이 빈번하게 발생하는 품사들을 확인하고 두가지 중 하나만 저장된 어절을 사전에서 제거하거나 분석 결과에 저장되지 않은 형태소의 결과를 추가하려고 하였다. 그러나 사전에 분석 결과를 임의로 추가 하는 것은 수동으로 작성한 정답만 저장하는 기분석 사전의 정의에 어긋나는 일이기 때문에 제외하였다.

전체 어절 중에서 혼동되는 문법 형태소 가운데 한가지만 결과에 포함한 어절을 조사하였다. 이것은 전체의 43.63%에 해당되어 모두 사전에서 제거할 경우 사전의 규모가 너무 작아질 수 있다. 또한 평가 파일 없이 혼동되는 문법 형태소를 찾아내는 것은 매우 어려운 일이고, 예외적인 경우가 많이 섞여있어 자동으로 어절을 제거하는 것이 어렵다. 따라서 동형이의 문법 형태소를 이용한 필터링 방법은 본 논문에 적용하지 않았다.

3.2 분석 결과 종류 수에 의한 필터링

기분석 사전에는 어절과 함께 어절의 형태소 분석 결과가 저장된다. 이때 어절마다 사전에 저장된 형태소 분석 결과의 종류는 서로 다르다. 단순하게 생각해 봤을 때 분석 결과의 종류가 많은 어절 보다 적은 어절이 불충분 분석 어절일 확률이 높다. 따라서 말뭉치에서 n가지 이상의 분석 결과로 출현한 어절만 이용해 기분석 사전을 구성하고자 하였다. 하지만 모든 어절을 분석 결과의 종류 수를 기준으로 필터링할 경우 중의성이 없는 어절이 모두 사전에서 제거되는 문제가 발생한다. 또한 Table 2에서 볼 수 있는 것처럼 사전의 크기가 너무 급격하게 감소하고, 동시에 사전의 적중률도 하락하기 때문에 분석 결과 종류 수를 기준으로 말뭉치를 필터링해 사전을 구성하는 것은 문제가 있다.

Table 2. Hit rate of analysis for the word-phrases classified with the number of analysis candidates

the number of analysis	the number of word-phrases (considering the frequency)	hit rate
1	800,340	92.75 %
2	34,115	37.02 %
3	3,013	15.32 %
4	786	8.20 %

3.3 출현 빈도 기준 필터링

어떤 어절이 말뭉치에 빈번하게 등장하였다는 것은 그 어절이 말뭉치에 등장한 결과로 자주 형태소 분석 된다는 것을 의미한다. 다시 말하면 해당 어절은 말뭉치에 등장하지 않은 형태소 분석의 의미로 사용될 확률이 낮다. Table 3은 말뭉치에서 출현한 빈도별로 어절을 저장한 사전에 포함된 불충분 분석 어절의 비율을 측정한 결과로, 말뭉치에 출현한 빈도가 높은 어절일수록 불충분 분석 어절일 확률이 낮음을 보여준다.

Table 3. Ratio of insufficient analysis in word-phrases classified with the number of word-phrase frequency

the number of word-phrase frequency	Ratio of insufficient analysis in word-phrases
1	6.05 %
2~3	2.76 %
4~7	0.32 %
8~15	0.08 %

이것을 근거로, 말뭉치에서 n번 이상 출현한 어절만 사전에 저장하여 사용하고자 한다. 필터링 기준이 높으면 높을수록 사전의 정답 제시율은 증가하겠지만 사전의 적중률이 떨어져 기분석 사전을 사용하는 의미가 없어진다. 따라서 형태소 분석기의 성능을 떨어뜨리지 않으면서, 적중률을 유지할 수 있는 적절한 필터링 기준을 제시할 필요가 있다.

위의 세가지 방법 이외에도 불충분 분석 어절의 길이가 충분 분석 어절의 길이보다 특별히 길거나 짧을 수 있다는 가정에 의해서 충분 분석 어절과 불충분 분석 어절의 길이를 측정해 보았다. 충분 분석 어절의 길이는 평균 3.54 어절, 불충분 분석 어절의 길이는 평균 3.48 어절로, 어절의 길이를 특징으로 보고 사전을 필터링하는 것은 어려웠다.

4. 기분석 사전의 성능 측정

사전 구성시 사용되는 말뭉치의 크기에 따라 변화하는 사전의 성능을 측정하기 위해 말뭉치의 크기를 균등하게 분리해 사용할 필요가 있다. 이에 실제 어절 분포와 유사하게 말뭉치를 분리하기 위해 어절을 문장단위로 묶어 100개의 파일에 균등하게 분배하였다. 이때 실험에 사용한 말뭉치는 3장에서 사용한 말뭉치와 동일하다. 이렇게 분리된 100개의 파일을 9:1로 나누어(이때 하나의 파일은 약 7만 어절로 구성된다.) 90개는 사전으로, 10개는 평가 파일로 사용하였으며, 사전으로 사용하는 파일과 평가용으로 사용하는 파일을 서로 다르게 하여 10-fold test와 유사한 방식으로 10번의 성능 측정을 하였다. 말뭉치의 크기는 사전으로 사용되는 90개의 파일을 하나씩 합쳐가면서 사용해 변화시켰고, 숫자 n을 필터링 기준으로 하여, n번 이상 출현한 어절만 추출해 사전으로 구성하였다.

4.1 기분석 사전의 적중률 측정

실제 형태소 분석기에 입력된 어절이 사전에 저장되어 있을 확률을 적중률이라 한다[2,6]. 본 논문에서는 사전구성에 사용된 파일과 겹치지 않는 10개의 평가용 파일을 실제 입력어절의 모집단으로 하여 사전의 적중률을 측정하였다.

$$\text{적중률} = \frac{\text{기분석사전에서 탐색된 어절의 수}}{\text{전체 어절의 수}} \quad (1)$$

기분석 사전의 적중률은 사전 구성시 사용한 말뭉치의 크기와 필터링 기준에 따라 Fig. 2의 그래프와 같은 양상으로 변화한다.

파일의 크기가 증가할수록 파일을 추가할 때 새로운 어절이 등장할 확률이 낮아지고 이미 사전에 저장된 어절이 다시 등장할 확률이 높아진다. 이렇게 재등장한 어절이 새로운 어절보다 빈번하게 사용되는 어절일 확률이 높다. 따라서 사전 구성에 사용된 파일의 수가 적을 때는 파일이 합쳐질 때마다 적중률이 급격하게 증가하고, 어절의 필터링 기준이 올라갈 때마다 적중률이 감소하는 정도가 매우 크다. 하지만 파일의 수가 많아질수록 파일이 합쳐질 때마다 적중률이 증가하는 정도와 필터링 기준 상향조정에 따른 적중률 감소폭이 줄어든다. 이것은 Heap's law가 영어 단어 뿐만 아니라 한국어 어절에도 유사하게 적용되는 것을 보여준다.

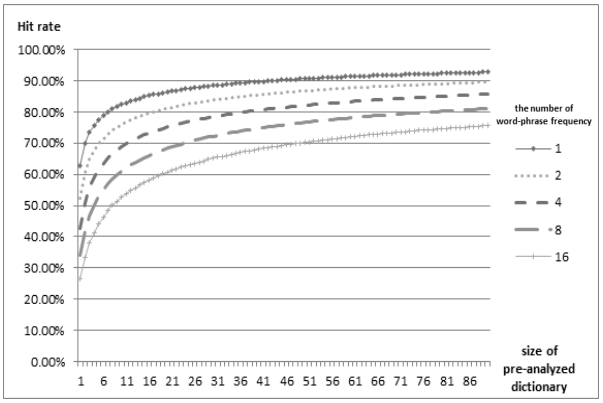

Fig. 2. Hit rate of pre-analyzed dictionary

수치를 자세히 살펴보기 위해 합쳐진 파일의 수가 1, 23, 45, 67, 90이 되는 5개의 지점을 잡아 Table 4를 구성하였다.

합쳐진 파일의 수가 1개일 때 모든 어절을 이용한 사전과 16번 이상 출현한 어절로만 구성한 사전의 적중률을 35%p 이상 차이가 나지만 합쳐진 파일 수가 90개일 때는 20%p 미만으로 그 폭이 크게 줄어들었다. 또한, 파일 수가 1개에서 23개로 증가 할 때 사전의 적중률은 20%p 이상 증가했지만 67개에서 90개로 증가할 때는 3%p 미만으로 증가하는 것을 확인할 수 있다.

4.2 기분석 사전의 충분분석률 측정

기분석 사전을 이용해 형태소 분석을 했을 때 생성된 후보에 정답이 없는 경우는 해당 어절이 불충분 분석 어절인 경우이다. 따라서 기분석 사전으로 형태소 분석이 가능한 어절 중 올바른 분석이 되는 경우의 비율을 충분 분석률이라 하고 식 (2)와 같이 정의하였다.

Table 4. Hit rate of pre-analyzed dictionary in accordance with corpus size and the number of word-phrase frequences

the number of word-phrase frequency \ the number of files	1	23	45	67	90
1	62.69%	87.15%	90.21%	91.74%	92.75%
2	52.42%	82.13%	86.15%	88.21%	89.58%
4	42.63%	76.35%	81.37%	83.95%	85.70%
8	33.92%	69.70%	75.79%	78.92%	81.09%
16	26.37%	62.22%	69.24%	73.02%	75.59%

$$\text{충분 분석률} = \frac{\text{정답 제시율}}{\text{적중률}} = \frac{\text{생성 후보 중 정답이 포함된 어절 수}}{\text{사전에서 탐색 된 어절 수}} \quad (2)$$

$$\text{정답 제시율} = \frac{\text{생성 후보 중 정답이 포함된 어절 수}}{\text{전체 어절 수}} \quad (3)$$

사전의 충분 분석률은 사전 구성시 사용한 말뭉치의 크기와 출현 빈도 기준에 따라 Fig. 3의 그래프와 같은 양상으로 변화한다.

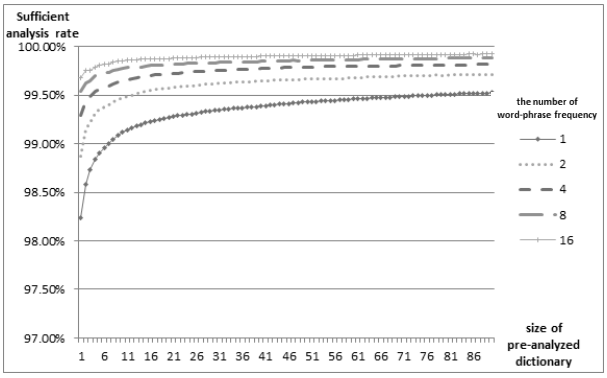

Fig. 3. Sufficient analysis rate of pre-analyzed dictionary

충분 분석률은 사전을 구성하는 말뭉치의 크기가 클수록, 말뭉치를 필터링 하는 빈도 기준이 높을수록 증가한다. 적중률과 같이 충분 분석률도 말뭉치의 크기가 커질수록 증가하는 정도가 감소하는데 적중률을 측정할 때, 한국어 어절에 Heap's law를 적용할 수 있었던 것 처럼, 형태소 분석 결과를 하나의 새로운 단어처럼 보고 Heap's law를 적용 할 수 있다. 적중률 측정 시와 같이 합쳐진 파일 수의 5개 지점을 잡아 Table 5를 구성해 자세한 수치를 살펴보았다.

5. 통합 시스템의 평가

본 논문에서 통합 시스템이란 기분석 사전과 형태소 분석기를 동시에 사용하는 시스템을 말한다. 즉, 입력 어절에 대해 우선 기분석 사전을 찾아보고 적중하면 그 결과를 출력으로 내고, 그렇지 않으면 형태소 분석기를 통해 결과를 출력하는 시스템이다.

통합 시스템의 형태소 분석기로써 통계 기반의 3단계 형태소 분석기인 SMA[13]를 사용하였다. SMA는 말뭉치에서 자동으로 구축되는 형태소 분석기이므로 기분석 사전의 말뭉치를 공유할 수 있다. 즉, 실험에 사용한 SMA는 기분석 사전 구축에 사용한 파일 90개를 그대로 학습하여 사용하였다.

이 시스템에서, 기분석 사전만 사용할 경우, 미적중된 어절은 정답 제시율이 0%이다. 만약 미적중된 어절에 대해서 추가로 형태소 분석기를 사용하면 정답 제시율은 최소한 0%보다 크거나 같을 것이다. 적중된 어절에 대해서는 기분석 사전의 정답 제시율과 형태소 분석기의 정답 제시율이 다르므로, 둘 중 정답 제시율이 높은 결과를 사용하는 것이 유리하다.

현재의 통합시스템에서 기분석 사전의 충분분석률이 어느 정도 높아야 SMA와 같이 사용하는 것이 효과적인지를 알기위해 평가하였다. 이를 위해 평가파일에 나타난 어절 중 기분석 사전에 적중되는 어절만을 대상으로 SMA의 성능을 측정하였다. 그 결과는 Fig. 4와 같다.

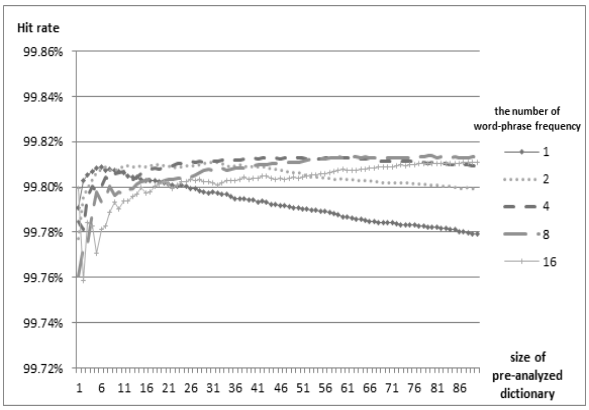

Fig. 4. Recall of SMA for the registered word-phrases

그래프를 살펴보면 사전에 적중한 어절에 대한 SMA의 성능이 최대 99.82%를 넘지 못하는 것을 알 수 있다. 따라

Table 5. Sufficient analysis rate of pre-analyzed dictionary in accordance with corpus size and the number of word-phrase frequency

the number of files / the number of word-phrase frequency	1	23	45	67	90
1	98.23%	99.29%	99.41%	99.48%	99.53%
2	98.87%	99.59%	99.65%	99.69%	99.72%
4	99.29%	99.73%	99.78%	99.80%	99.82%
8	99.54%	99.82%	99.85%	99.87%	99.88%
16	99.68%	99.88%	99.91%	99.91%	99.92%

Table 6. Ratio of unknown words

	total word-phrase in test file	hitted word-phrases	non-hitted word-phrases
the number of word-phrases	757,340	702,432	54,908
the number of unknown words	5,354	232	5,122
ratio of unknown words	0.71%	0.03%	9.33%

Table 7. Recall of integrated system (SMA with pre-analyzed dictionary)

the number of files / the number of word-phrase frequency	1	23	45	67	90
1	98.08%	98.62%	98.71%	98.78%	98.83%
2	98.59%	98.88%	98.93%	98.96%	98.99%
4	98.85%	99.00%	99.03%	99.05%	99.07%
8	98.97%	99.07%	99.09%	99.11%	99.12%
16	99.01%	99.11%	99.13%	99.14%	99.14%
32	98.97%	99.13%	99.14%	99.15%	99.16%
64	98.88%	99.12%	99.15%	99.15%	99.16%
128	98.67%	99.10%	99.12%	99.13%	99.13%

서 기분석 사전의 충분 분석률이 99.82%이상이 되어야만 통합 시스템의 성능을 떨어뜨리지 않는다는 것을 알 수 있다.

사전에 적중한 어절에 대한 SMA의 정답 제시율이 99.82%로 높은 이유를 분석해 보면, 통합 시스템의 특성상 기분석 사전에 저장된 어절의 분석결과가 대부분 SMA의 형태소 사전에 저장되어 있기 때문에 나타나는 현상이다. SMA는 미등록어가 있을 경우, 그 오류률이 높지만, 등록된 형태소들에 대해서는 정답 제시율이 높은 편이다. 이를 확인하기 위해 실제로 실험에 사용한 평가용 말뭉치와 사전용(학습용) 말뭉치를 이용해 이를 확인해 보았다. 그 결과는 Table 6과 같다. 즉, 적중된 어절이 적중되지 않은 어절에 비해 훨씬 적은 미등록어를 포함하고 있다.

사전을 구성할 때 사용한 말뭉치 크기와 출현 빈도 기준에 따른 충분 분석률을 측정한 4장의 Table 5를 살펴보면 충분 분석률이 99.82%이상인 기분석 사전을 구성하기 위해서는 약 160만 어절(23개의 파일) 이상의 말뭉치를 이용해 8번 이상 출현한 어절로 기분석 사전을 구성해야 한다. 이때 적중률은 69.70%이다. 사전을 구성할 때 사용하는 말뭉치의 크기가 작으면 출현 빈도 필터링 기준을 올려 충분 분석률을 향상 시킬 수 있지만 적중률이 급격하게 감소한다. 그리고 필터링 빈도 기준이 이보다 낮으면, 말뭉치의 크기를 증가 시켜 충분 분석률을 향상 시킬 수 있으나 말뭉치를 구성하는데는 많은 비용이 필요하다. 따라서 말뭉치의 크기에 따라 통합 시스템을 구성할 때 가장 적절한 출현 빈도 필터링 기준을 확인할 필요가 있다. 이에 SMA와 기분석 사전을 결합한 통합 시스템의 성능을 측정하였다(Fig. 5, Table 7).

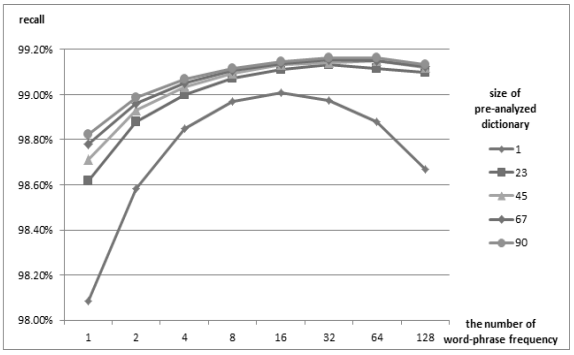

Fig. 5. Recall of integrated system (SMA with pre-analyzed dictionary)

7만 5천여 어절(파일 1개)만 이용해서 기분석 사전을 구성한 경우 출현 빈도 필터링 기준을 아무리 올려도 통합 시스템의 성능이 SMA성능을 넘지 못한다. 하지만 160만여 어절(파일 23개)이상 크기의 말뭉치를 이용해 기분석 사전을 구성한 경우, 출현 빈도 필터링 기준을 올리면, 통합 시스템의 성능이 SMA의 성능 보다 나아지는 것을 확인 할 수 있다. 그러나 출현 빈도 필터링 기준을 계속해서 올리면 어느 순간 통합 시스템의 성능이 다시 떨어지는 것을 확인 할 수 있다. 이것은 사전의 충분 분석률이 증가하는 정도보다 사전의 적중률이 떨어지는 정도가 커지기 때문인데, 일정 수준 이상의 성능을 보장하는 기분석 사전에서 출현빈도 기준을 올려 계속 어절을 필터링 하면, 사전을 이용한 분석을 할 때 더 정확한 분석을 할수 있는 어절이 SMA에서 분석되어 실패하는 경우가 발생하기 때문이다.

이것으로, 160만 여 어절로 사전을 구성 할때는 필터링 빈도 기준을 32로 하는 것이, 630만여 어절로 사전을 구성 할 때는 필터링 빈도 기준을 64로 하는것이 가장 최적임을 파악 할 수 있었다. 또한 말뭉치의 크기가 커질수록 통합 시스템의 성능이 최적이 되는 출현 빈도 필터링 기준이 높아지는 것을 확인할 수 있었다.

6. 결 론

기분석 사전은 형태소 분석기의 처리 속도와 정확도를 높이고, 과분석을 줄이기 위해 사용한다. 그런데 사전에 불충분 분석 어절이 존재하여 오히려 형태소 분석기의 정답 제시율을 떨어뜨리는 경우가 발생한다. 본 논문에서는 형태소 분석기의 정확도를 높이기 위한 기분석 사전 구성 방법을 제시하고 실제 형태소 분석기와 통합하여 그 성능을 측정하고 평가하였다. 그리고 통계기반 3단계 형태소 분석기인 SMA와 사전을 결합하여 통합 시스템의 성능을 최적화 하는 기분석 사전 구성 기준을 확인 하였다.

불충분 분석 어절을 사전에서 제거할 때에 논문에서 제시한 빈도 기준의 방법 외에도 다른 방법이 존재할 수 있다. 특히 동형이의 문법 형태소를 더 조사한다면 불충분 분석 어절을 효율적으로 처리할 수 있는 방법을 찾을 수 있을 것이라 예상한다.

참 고 문 헌

[1] S. S. Kang, “Korean Morphological Analysis and Information Retrieval,” *Hongrung Publisher*, 2002.

[2] J. H. Kim, C. Y. Ok, “Korean Morphological Analysis using Inflected-Word-Dictionary,” *Proceedings of Spring Conference on KIISE*, Vol.21, No.1, pp.813-816, 1994.

[3] S. S. Kang, “Encoding of Morphological Analysis Result and Eojeol Dictionary Construction,” *Proceeding of the 16th Hangul and Korean Information Processing conference on KIISE*, Vol.16, No.1, pp.112-117, 2004.

[4] D. Lee, “Probabilistic Models for Korean Morphological Analysis and Part-of-Speech Tagging,” Ph.D. dissertation, University of Korea at Computer Science and Engineering, Korea, 2005.

[5] J. C. Shin, C. Y. Ock, “A Korean Morphological Analyzer using a Pre-analyzed Partial Word-phrase Dictionary,” *The KIISE Transactions: Software and Application*, Vol.39, No.5, pp. 415-424, 2012.

[6] S. S. Kang, “Extracting High-Frequency Optimal Korean Word Set by Word Frequency Statistics,” *Proceeding of the 13th Hangul and Korean Information Processing conference on KIISE*, pp.85-88, 2001.

[7] S. H. Yang, Y. S. Kim, “A High-Speed Korean Morphological Analysis Method based on Pre-Analyzed Partial Words,” *The KIISE Transactions: Software and Applications*, Vol.27, No.3, pp.290-301, 2003.

[8] S. S. Kang, Y. T. Kim, “A Computational Analysis Model of Prefinal Endings for Korean Morphological Analyzer,” *The KIISE Transations*, Vol.18, No.5, 1991.

[9] J. P. Hong, J. W. Cha, “A New Korean Morphological Analyzer using Eojeol Pattern Dictionary,” *Proceeding of the Korea Computer Congress on KIISE*, Vol.35, No.1(C), pp.279-284, 2008.

[10] H. S. Lim, H. Lee, H. C. Rim “A Method of Analyzing Word Ambiguity in Korean Morphological Analysis,” *Proceeding of Spring Conference on KIISE*, Vol.20, No.1, pp.779-776, 1993.

[11] J. S. Nam, K. S. Choi, “Disambiguation Method based on a Lexicon of Typographical Units,” *Proceeding of the 9th Hangul and Korean Information Processing conference on KIISE*, pp.75-82, 1997.

[12] J. Y. Kee, K. Y. Lee, H. W. Kim “Disambiguation in Morphological Analysis Using Word Rules,” *Proceeding of autumn Conference on KIISE*, Vol.24, No.2(Ⅱ), pp.215-218, 1997.

[13] The National Institute of the Korean Language, 21st Century Sejong Project Final Result - 2011.12. Revised Edition, 2011.

[14] J. S. Lee, “Three-Step Probabilistic Model for Korean Morphological Analysis,” *The KIISE Transactions: Software and Application*, Vol.38, No.5, pp.257-268, 2011.

곽 수 정

e-mail : crystalk@cbnu.ac.kr

20011년 충북대학교 컴퓨터교육과(학사)

2012년~현 재 충북대학교 정보산업 공학협동과정 석사과정

관심분야 : 자연 언어 처리, 정보 검색

김 보 겸

e-mail : bogyum@cbnu.ac.kr

2007년 충북대학교 컴퓨터교육과(학사)

2011년 충북대학교 정보컴퓨터교육과 (석사)

2011년~현 재 충북대학교 디지털정보 융합학과 박사과정

관심분야 : 자연 언어 처리, 정보 검색

이 재 성

e-mail : jasonlee@cbnu.ac.kr

1983년 서울대학교 컴퓨터공학과(학사)
1985년 KAIST 전산학과(석사)
1999년 KAIST 전산학과(박사)
1985년~1988년 큐닉스 컴퓨터 과장
1988년~1993년 미국 및 한국 마이크로소프트 개발부 차장
1999년~2000년 ETRI 컴퓨터소프트웨어기술연구소 팀장
2005년~2006년 (미) 아리조나 대학 방문 교수
2000년~2011년 충북대학교 컴퓨터교육과 교수
2011년~현 재 충북대학교 디지털정보융합학과 교수
관심분야 : 자연 언어 처리, 정보 검색, 컴퓨터 교육 등