말뭉치를 이용한 한국어 단어 개수 추정

김 성 기† · 한 근 식††

요 약

한 언어에서 사용되는 단어의 총 개수를 추정하는 것은 매우 어려운 작업이다. 최근 한 언어를 대표하는 것으로 생각되는 원문, 발화, 또는 기타 표본들의 뭉치인 말뭉치가 대규모로 구축됨으로 말뭉치를 기반으로 하여 한 언어의 총 단어 개수를 추정할 수 있게 되었다. 본 논문에서는 한국어 말뭉치에 나타난 단어를 기반으로 한국어 단어의 총 개수를 추정하는 방법을 제시하고 한국어 단어의 총 개수를 추정한다. 이와 더불어 한국어에서 가장 많은 수의 고유명사를 차지하는 한국사람 이름의 총 개수도 함께 추정한다. 단어 개수와 이름 개수의 추정방법은 빈도를 이용한 일반화된 선형모형을 적용하였다. 1000만 어절의 말뭉치를 이용하여 한국어의 총 단어를 추정한 결과 1,062,392개로 추정되었으며 한국사람 이름의 개수는 1,493,003개로 추정되었다.

Estimating the Number of Korean Words Based on Corpus

Sung-ki Kim† · Geun-shik Han††

ABSTRACT

It is very hard to estimate the number of total words in a language. Recently large corpus which is the body of written, spoken or other material and which is thought as the representative of a language is under construction. So, it is possible to estimate the number of words in a language based on the corpus. In this paper we propose the method for estimating the number of Korean words using Korean corpus and estimate the number of words. We also estimate the number of Korean names which occupy the large part of proper nouns. To estimate the number of total different Korean words and names we applied a generalized linear estimation method. 1,062,392 is the number of estimated Korean words using the corpus of 10 million phrases and 1,493,003 is the estimated number of Korean names.

1. 서 론

알려지지 않은 새로운 종(species)의 수를 추정하는 문제가 생태학 연구분야에서 연구되어 왔다. 이와 유사한 연구로 Efron과 Thisted[8, 9], McNeil[12] 등은 특정 저자가 알고는 있지만 저서에서 이용되지 않은 단어들을 알려지지 않은 새로운 종에 연결하여 저자가 실제로 알고 있을 것으로 생각되는 단어의 총수를 추정하는 연구를 하였다. 이러한 연구는 단어의 사용빈도(frequency)를 이용하는 방법과 베이지안(Bayesian) 접근법을 이용하는 방법 등이 주류를 이루고 있다. [8]에서는 세익스피어의 작품에 나타난 단어의 빈도를 이용하여 세익스피어가 알고 있는 총 단어의 개수를 최소한 66,534개라고 추정하였다.

국내의 한국어 어휘에 대한 연구로서 국어 어휘의 통계적 빈도수[6], 고빈도 단어 리스트[3] 등의 연구가 이루어졌다. 특히 한국어는 영어와 같은 서구어와는 달리 교착어의 특징이 강한 언어로서, 용언의 변화형이 매우 복잡하고 상과 서법에 관련된 문법요소가 개입하

※ 이 논문은 1997년도 한신대학교 응용과학연구소 연구비 지원에 의하여 연구되었음.
† 정 회 원 : 한신대학교 전산학과 교수
†† 정 회 원 : 한신대학교 전산통계학과 교수
논문접수 : 1997년 11월 28일, 심사완료 : 1998년 5월 7일

고 있으며 복합동사와 보조 용언의 쓰임이 매우 활발하고 명사에 있어서도 복합 명사와 조사의 쓰임이 매우 복잡하므로 어휘 조사는 조사의 목적이나 조사 단위의 선정에 따라 결과가 매우 달라질 수 있다[2]. 대규모 한국어 말뭉치가 구축됨으로 인하여 한국어 어휘 조사 및 어휘에 대한 연구는 다양한 각도에서 심도 깊게 이루어질 것으로 예상된다. 그러나 아직 한국어 어휘량 또는 단어의 총 개수에 대한 연구는 필자가 아는 한 시도된 바가 없다.

최근 정보처리기술의 급속한 발전에 따라 방대한 양의 한국어 텍스트 수집과 저장이 용이해졌으며 이를 토대로 한국어의 언어 현상이 국어학, 국어공학 분야에서 실증적으로 연구 개발될 수 있는 기반이 마련되고 있다[7]. 한 언어를 대표하는 것으로 생각되는 원문, 발화, 또는 기타 표본들의 뭉치를 말뭉치(corpus)[5]라 하며 현재 1000만 어절 이상의 대규모 말뭉치가 국내에서도 연세대학교 한국어 사전편찬실, 고려대학교 민족문화연구소, 한국과학기술원 인공지능연구센터 등에서 구축되고 있다. 말뭉치를 이용하여 기본 어휘자료의 각종 통계치 조사, 통사 연구, 사전편찬, 문장 분석 등의 연구가 활발히 진행되고 있다[4, 6].

한국어의 전체 어휘량 또는 단어수가 얼마인가는 매우 흥미로운 물음이지만 이는 매우 어려운 물음이라고 할 수 있다. 한 언어의 단어는 부단히 생성되고 소멸되고 있으므로 정확히 그 개수를 셀 수는 없다. 최근 출간된 국어대사전에는 약 50만 단어를 수록하고 있으며 한국사람의 이름, 지명 등의 고유명사와 외국어 등을 고려하면 그 수효는 매우 많다라고 예측할 수 있을 것이다. 본 연구에서는 한국어 말뭉치를 구성하는 단어들을 분석하여 한국어에서 사용되리라고 예상되는 총 단어의 개수를 추정한다. 물론 조사대상의 말뭉치가 한국어의 모든 언어현상을 완전히 대표하지 못하고, 말뭉치에서 완벽한 단어분석이 어려우며 추정방법상의 오차 등으로 인하여 정확한 총 단어 개수의 추정은 어려울 것이지만 대규모 말뭉치를 이용할 경우 어느 정도의 신뢰도를 가지는 총단어 개수의 추정이 가능할 것이며 이를 위한 추정방법을 제시한다.

한국어 단어의 총 개수 추정에 대한 연구는 여러 면에서 중요한 의미를 가진다고 할 수 있다. 첫째, 한국어의 총 단어수를 추정함으로 전체적인 한국어 어휘량의 범위를 예측할 수 있게 한다. 둘째, 한국어 단어는 가변 길이의 음절로 구성되는데 보다 효율적인 정보처리를 위하여 한국어 단어 코드를 작성할 경우 단어의 총 개수는 단어 코드에 큰 영향을 미치게 된다. 셋째, 한국어 단어 개수 추정에 대한 방법론이 개발될 것이다. 마지막으로 한국어 단어 개수 추정은 말뭉치의 구축, 형태소 분석, 의미 분석 등 한국어 처리기술에 의존하므로 한국어 처리기술의 발전을 도모할 수 있을 것이다.

한국어 단어 개수의 추정은 총 N개로 구성된 모집단을 이용하여 말뭉치를 형성할 때 임의의 단어는 포아송 절차(Poisson process)에 따라 말뭉치에 나타나며 출현빈도는 이항분포를 따른다는 가정 하에서 출발하였다. 말뭉치에 나타난 단어의 출현빈도를 단어의 기대값으로 이용하는 과정에서 기대값의 수렴을 빠르게 하기 위하여 일반화된 선형모형(general linear model)을 설정하였으며 이 과정에서 포아송 분포의 이항분포 근사를 이용하였다.

출현빈도의 관측은 한국과학기술원 전산학과 인공지능센터에서 개발한 1000만 어절의 말뭉치를 이용하였다. 이 말뭉치는 "대한민국 국어정보베이스" 평가판 0.1로서 시스템 공학연구소 자연어 처리부의 협력 하에서 구축되었으며 1000만 어절의 말뭉치는 문학, 과학, 시사 등 여러 분야의 텍스트에서 선별한 가공되지 않은 원문으로 구성되어 있다.

출현빈도를 이용한 일반화된 선형모형으로 단어 개수를 추정한 결과를 보면, 말뭉치에서 추출된 총 단어 8,872,590개중에서 서로 다른 단어의 총 개수는 285,629개이었으며 이를 토대로 새로이 추정된 단어의 개수는 776,763개이었다. 그 결과 말뭉치에 근거한 한국어 단어의 총개수는 285,629+776,763 = 1,062,392개로 추정되었다. 또한 말뭉치 단어를 다시 9개의 그룹으로 분류하여 각 그룹에 대하여 새로운 단어를 추정한 결과 총 780,510개가 새로이 추정되었으며 이는 전체 단어의 추정치와 비슷한 결과를 보였다. 한편 말뭉치에 근거한 단어의 추정과 별도로 고유명사의 개수를 추정하기 위하여 고유명사중 가장 많을 것으로 예상되는 한국사람 이름의 개수를 대규모 인명자료를 이용하여 별도로 추정하였다. 인명자료의 이름의 총개수는 3,653,772개이었으며 서로 다른 이름의 개수 456,453개에서부터 1,036,550개의 이름이 새로이 추정되어 한국사람 이름의 총 개수는 456,453+1,036,550 = 1,493,003개이었다. 이름 이외의 지명, 상호명, 관직명 등의 여러 고유명사를 고려한다면 한국어에서 사용되는

총 단어의 개수는 약 300만개 이상이 됨을 예상할 수 있으며 이는 바로 한국어의 어휘량이 될 것이다.

이러한 추정치의 오차는 오일러(Euler) 변환식의 수렴속도에 영향을 받는데 말뭉치 단어나 인명자료의 이름의 경우 그 수렴속도가 빠르기 때문에 총 추정치에 대한 표준편차가 양호하다고 볼 수 있다. 말뭉치 단어의 경우 추정치와 표준편차는 776,763 ± 21,242로서, 추정치 776,763에 대한 표준편차가 21,242개로 표준편차에 대한 추정치의 비율이 2.7%이므로 추정치에 대한 신뢰도가 높다고 할 수 있겠다. 이름에 대한 추정치와 표준편차는 1,036,550 ± 57,196이며 추정치 1,036,550에 대한 표준편차가 57,196개로 표준편차에 대한 추정치의 비율이 5.5%로 신뢰도가 높다고 할 수 있겠다.

본 논문의 구성은 다음과 같다. 2장에서는 말뭉치에서 단어를 추출하여 분석하는 방법과 분석결과를 소개하며, 3장에서는 출현빈도를 이용한 일반화된 선형모형을 설정하여 단어와 이름의 개수를 추정하는 방법을 설명한다. 4장에서는 결론과 앞으로의 연구방향을 제시한다.

2. 말뭉치의 단어 분석방법 및 분석결과

본 논문에서는 한국과학기술원 전산학과 인공지능연구센타에서 개발한 1000만 어절의 말뭉치를 기반으로 하여 우선 말뭉치에 포함된 문장들에 대하여 형태소 분석을 실시하여 단어를 추출하였다. 형태소 분석에 사용된 한국어 형태소 분석기[1]는 한성대학교 한국어처리연구실에서 개발한 HAM을 이용하였다. HAM은 문장의 각 어절을 독립적으로 분석하여 그 어절에서 가능한 모든 형태소를 생성하는 형태소 분석기로서 어절의 의미를 고려하지 않고 형태적으로 어절을 분석하며 형태소 분석을 위하여 단어사전을 이용한다.

본 논문에서는 말뭉치에서 추출된 단어를 9개 그룹으로 분류하였다. 9개의 단어분류는 사용된 형태소 분석기 HAM이 적용한 단어 분류에 근거한 것이며 국어어휘론의 일반적 품사분류와는 다른 분류이다. 분석된 단어는 형태소 분석기의 단어사전에 등록된 등록체언, 등록된 체언들로 구성되는 등록복합명사, 단어사전에 미등록된 미등록체언, 단어사전에 등록된 등록용언, 단어사전에 미등록된 미등록용언, 보조용언, 부사, 조사, 그리고 기타(관형사, 감탄사 등)의 그룹으로 분류된다. 추출된 단어를 9개 그룹으로 분류하여 각 그룹별 개수를 추정하는 것은 층화추출의 효과를 반영할 수 있으며 이로 인해 추출된 단어를 하나의 그룹으로 처리하여 개수를 추정할 때 발생하는 과소추정성(underestimation tendency)를 배제하기 위함이었다.

한국어 문장에서 단어를 분석하고 추출할 때 단어분석 및 추출기준에 따라 그 결과는 매우 달라질 수 있다. 본 논문에서 일반적인 단어 분석 및 추출기준을 적용하는 한편 특정적인 중요한 기준은 다음과 같다. 첫째, 같은 분류와 같은 형태를 가지는 단어는 하나의 단어로 간주된다. 예를 들어 '우리'라는 단어는 대명사 '나'의 복수, 동물을 가두는 울타리 등의 여러 의미를 갖지만 본 논문의 분류상 등록체언에 포함되며 형태가 같으므로 본 연구에서는 하나의 단어로 간주한다. 의미를 고려하는 단어분석은 자연어 처리에서도 아직 완전히 이루어지지 않고 있으므로 차후에 계속되어야 할 과제이다. 둘째, 형태소 분석의 결과 여러 형태의 단어로 분석되는 모호성이 있는 어절은 분석대상에서 제외되었다. '나는'이라는 어절은 형태소 분석기에 의하여 '나'(등록체언)+는(조사), 날(등록용언)+'은'(어미) 또는 '나'(등록용언)+'는'(어미) 등 여러 형태로 분석되므로 분석대상에서 제외한다. 셋째, 복수형 단어는 단수형 단어로 분석하여 복수형 단어를 새로운 단어로 취급하지 않는다. 예를 들어 '그들', '책들' 등은 단수형인 '그'(등록체언), '책'(등록체언)으로 분류되어진다. 넷째, '피곤하다', '분석하다' 등 등록체언+'하다'가 결합된 어절은 등록체언으로 분석하였다.

형태소 분석기 HAM을 이용하여 1000만 어절 말뭉치에서 단어를 분석한 결과 추출된 단어는 8,872,590개이었다. 한국어의 한 어절에는 하나 이상의 단어가 포함되므로 1000만 어절인 경우 정확히 분석할 경우 총 추출단어는 1000만개가 넘을 것이지만, 여러 경우로 분석된 모호성이 있는 어절은 분석대상에서 제외시켰으며 한자어, 영어 그리고 특수문자 등이 제외되어 총 추출 단어는 1000만개보다 작았다. 8,872,590개의 단어 중에서 서로 다른 단어의 개수는 285,629개이었으며 한번 나타난 단어가 164,855개, 두번 나타난 단어가 37,593개이었다.

한편 말뭉치에는 성명, 지명 등의 고유명사가 많이 포함되지 않는다. 본 연구에서는 말뭉치에 포함되는 일반 단어 외에 고유명사의 개수를 별도로 추정하기 위하

〈표 1〉 분석된 단어 및 이름의 총개수와 빈도
〈Table 1〉 Total number and frequency of analyzed words and names

		총출현개수	서로다른개수	1회출현개수	2회출현개수
말뭉치단어	등록체언	3,401,024	35,074	5,217	3,041
	등록복합명사	168,369	56,272	36,000	8,014
	미등록체언	671,098	183,091	120,192	25,461
	등록용언	1,666,942	6,259	1,293	561
	미등록용언	3,167	1,758	1,398	199
	보조용언	74,353	187	98	30
	부사	12,714	741	307	87
	조사	2,576,482	356	27	20
	기타	299,441	1,891	323	180
	전체 단어	8,872,590	285,629	164,855	37,593
이름		3,653,772	456,453	199,768	74,628

여 고유명사중 가장 많은 개수를 차지할 것으로 예상되는 한국사람 이름의 총 개수를 추정하였다. 대규모 인명자료로부터 이름의 출현빈도를 구하여 총 단어 개수 추정과 같은 방법으로 총 이름 개수의 추정을 실시하였다. 인명자료의 이름 개수는 3,653,772개이었으며 서로 다른 이름의 개수는 456,453개이었고 한번 나타난 이름의 개수는 199,768개, 두번 나타난 이름의 개수는 74,628개이었다. 〈표 1〉에서는 말뭉치에서 분석된 단어의 분포와 인명자료에서 분석된 이름의 분포를 나타내고 있다.

3. 추정모형

본 장에서는 출현빈도를 이용한 모집단의 개수 추정을 위한 일반화된 선형모형을 제시하고 이를 기반으로 하여 한국어 말뭉치에서의 단어 개수와 인명자료에서의 이름 개수를 추정하고 이들의 표준편차를 구한다. 먼저 Fisher[11]의 방법에 따라 기본모형에 대한 가정을 설명하면 한국어의 서로 다른 단어의 총 개수(모집단)를 N개라고 할 때 말뭉치에서 임의의 단어 s가 x_s번 이용되었다는 것을 확인할 수 있다. 여기에서 x_s는 임의의 단어 s의 출현빈도를 의미하며 0보다 큰 경우만 관측 가능하다. 모집단의 서로 다른 N개의 단어는 포아송 절차에 따라 말뭉치에 나타나며 서로 다른 N개의 단어가 말뭉치에 쓰여지는 절차는 말뭉치를 구축할 때 기대값 λ_s를 갖는다. 즉 x_s는 평균이 $\lambda_s (s=1,2,\cdots,N)$인 포아송 분포를 따른다.

$x_s(t)$를 구간 [-1, t]에서 단어 s의 출현빈도라 하면 이때 포아송절차에 대한 가정은 첫째, $x_s(t)$가 평균이 $\lambda_s(1+t)$인 포아송 분포를 따르며; 둘째, $x_s(t)$가 주어졌을 때 출현빈도 x는 모수(parameter)가 $\frac{1}{1+t}$인 이항분포를 따른다는 것이다. 즉 $x \sim Bin\left(x_s(t), \frac{1}{1+t}\right)$이다. 이 연구에서 t는 주어진 말뭉치가 아닌 다른 곳에서 새로이 발견된 단어의 총개수를 이미 발견된 말뭉치 단어의 총개수로 나눈 값이 된다. 즉 말뭉치 단어의 경우 $t=\frac{\text{새로이발견된단어의수}}{8,872,590}$가 된다. 첫번째 가정은 그다지 중요하지 않지만 두번째 가정은 빈도를 이용한 추정에서 상당히 중요하다. 한편 [-1, 0]은 연속함수를 이용하여 t=0일 때의 기대값을 구하기 위한 전체구간 [-1, t]의 형식적인 구간이다. $t \to 0$일 때는 새로이 발견된 단어들이 조선말기의 한글고어체 단어라든가 정보통신관련 전문용어 등으로 구성되어졌을 경우를 나타내며 이런 경우 이 연구의 결과는 신뢰성이 없을 것이라고 할 수 있다.

$G(\lambda)$를 $\lambda_1,\ldots,\lambda_S$의 경험적 누적분포(empirical cumulative distribution) 함수라 하자. 그리고 n_x를 [-1 , 0]에서 x번 관측된 단어의 수라고 하면

$$\eta_x = E(n_x) = S\int_0^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}\, dG(\lambda) \quad \text{———(1)}$$

이 되며, $\triangle(t)$를 (0, t]에서 x번 관측된 단어의 기대값이라 하자. 즉

$$\triangle(t) = S\int_0^{\infty} e^{-\lambda}(1-e^{-\lambda t})\, dG(\lambda) \quad \text{———(2)}$$

새로운 단어들의 집합을 발견하였을 때 현재의 총 단어수를 알고 있으므로 t를 계산할 수 있고 새로운 단어와 현재의 단어를 합성하여 만들어질 수 있는 단어의 수(기대값)가 $\triangle(t)$가 된다. 이 연구의 목적은 이를 추정하는 것이다. 식(2)를 이용하기 위해 $(1-e^{-\lambda t})$를 전개하면 다음과 같다.

$$1-e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \cdots \quad \text{———(3)}$$

위 식을 식(2)에 대입하고 식(1)과 비교하면

$$\triangle(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \cdots \quad \text{———(4)}$$

이 된다. 식(4)는 Good과 Toulmin[10]에도 나타나 있는데 만약 $\triangle(t)$가 수렴한다면 그것의 불편추정량(unbiased estimator)은 η_i를 표본의 값 n_i로 대체함으로서 다음과 같이 구할 수 있다.

$$\hat{\triangle}(t) = n_1 t - n_2 t^2 + n_3 t^3 - \cdots \quad \text{———(5)}$$

식(5)는 $t \le 1$ 경우에 유용하게 $\triangle(t)$를 추정하나 $t \gg 1$ 인 경우 $\triangle(t)$ 값의 진동폭이 심하다. Good과 Toulmin은 이러한 수열의 수렴을 빠르게 하도록 오일러 변환을 이용할 것을 권하고 있는데 이러한 변환을 통해서 진동이 큰 수열의 수렴을 빠르게 해줄 수 있다. $t=\frac{u}{3-u}$를 이용한 오일러 변환식은

$$\sum_{x=1}^{\infty}(-1)^{x+1}\eta_x t^x = \sum_{y=1}^{\infty}\zeta_y u^y$$

이 된다. 여기에서

$$\zeta_y = \sum_{x=1}^{y}\binom{y-1}{x-1}\frac{(-1)^{x+1}}{3^y}\eta_x = \frac{1}{3^y}\delta^y(\eta_1) \quad \text{———(6)}$$

이며 $\delta^y(\eta_1)$은 다음과 같다.

$$\delta^0(\eta_1) = \eta_1,\ \delta^1(\eta_1) = \eta_1 - \eta_2,\ \delta^2(\eta_1) = \eta_1 - 2\eta_2 + \eta_3$$

이제 오일러 변환식의 부분합(partial sum)을 $\triangle^{x_0}(t)$와 $\triangle^{x_0}(u)$로 표기하면 다음과 같이 표현된다.

$$\triangle^{x_0}(t) = \sum_{x=1}^{x_0}(-1)^{x+1}\eta_x t^x,$$

$$\triangle^{x_0}(u) = \sum_{y=1}^{x_0}\zeta_y u^y \quad \text{----------(7)}$$

식(7)의 부분합에 극한을 취하면 다음과 같이 된다.

$$\triangle(t) = \lim_{x_0\to\infty}\triangle^{x_0}(t),\quad \triangle(u) = \lim_{x_0\to\infty}\triangle^{x_0}(u)$$

만약 양쪽 극한이 존재하면 $\triangle(t) = \triangle(u)$가 된다. 그리고 $\triangle^{x_0}(u)$의 부분합은 $\triangle^{x_0}(t)$보다 극한에 빠르게 수렴한다. 이 연구에서는 식(6)을 추정하기 위해 η_x대신 n_x를 이용하였으며 $\triangle(t)$를 추정하기 위해 오일러 변환수열인

$$\triangle^{x_0}(u) = \sum_{y=1}^{x_0}\hat{\zeta}_y u^y,\quad u = \frac{3t}{1+t} \quad \text{———(8)}$$

을 이용하였다.

식(6)을 이용한 $\hat{\zeta}_y$의 값을 구한 결과는 〈표 2〉와 같다.

〈표 2〉는 식(6)을 이용하여 $\hat{\zeta}_y$을 계산한 것으로 전체단어의 경우 $y=10$에서 음의 값을 갖으며 이름과 등록복합명사의 경우 $y=11$에서 음의 값을 갖는다. η_x의 추정치 n_x를 이용하여 $\widehat{\triangle^{x_0}}(u)$를 계산할 때 x_0가 '0'의 값에 가장 근접한 t 값을 이용하기로 한다. 이제 $\triangle(t)$를 추정하기 위해 오일러 변환수열을 일반 선형추정량(general linear estimator)으로 변환하면 된다. 이를 위해 식(6)을 식(7)에 대입하면 $\triangle^{x_0}(u)$는 진동하는 수열 $\triangle^x(t)$의 평균과 같다. 여기서 x는 모수가 $\frac{1}{1+t}$인 이항분포를 따른다. 이제 식(8)은 η_x의 추정치로서 $\hat{\eta}_x = n_x$를 대입하면 식(9)와 같이 표현된다.

$$\hat{\triangle} = \sum_{x=1}^{\infty} h_x n_x \quad \text{———(9)}$$

여기서, $h_x = \begin{cases} (-1)^{x+1} t^x P(Z \ge x), & x = 1,\cdots,x_0 \\ 0, & x > x_0 \end{cases}$

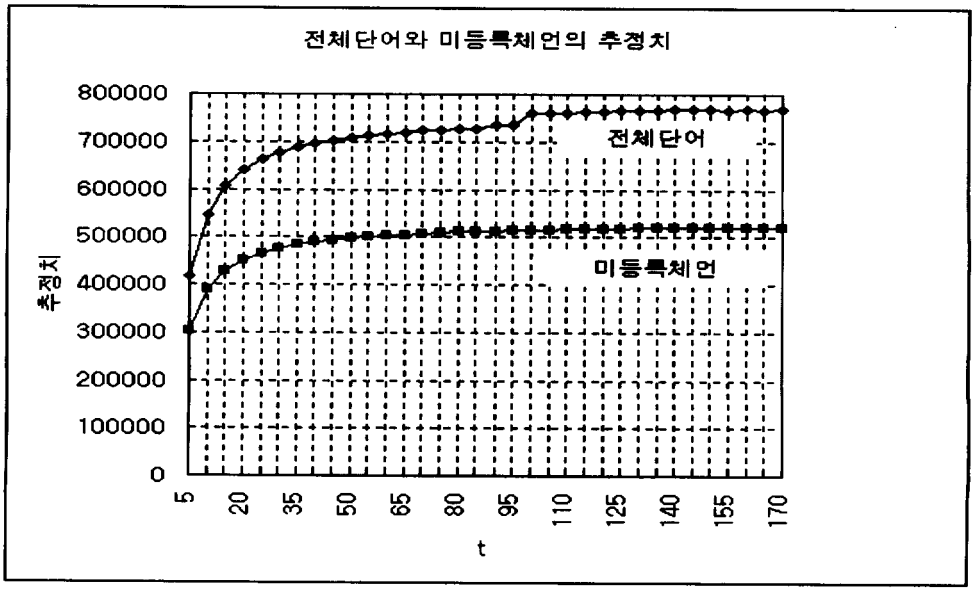
〈표 2〉 $\hat{\xi}_y$의 계산 결과

〈Table 2〉 Evaluation result of $\hat{\xi}_y$

y \ 항목		1	2	3	4	5	6	7	8	9	10	11
말뭉치 단어	등록체언	399	118	35.6	11.0	3.47	1.12	0.37	0.13	0.05	-0.03	
	등록복합명사	9328	2618	766	229	69.4	20.7	6.01	1.60	0.35	0.048	-0.006
	등록용언	244	56.4	15.7	5.7	2.6	1.3	0.7	0.35	0.15	0.059	-0.02
	미등록용언	22.6	5.4	1.29	0.2	0.03	-0.01					
	보조용언	399.6	118	35.6	11.0	3.47	1.12	-0.04				
	부 사	132	19.4	4.8	0.9	-0.04						
	조 사	2.3	-0.3									
	기 타	47.6	9.4	2.2	0.7	0.39	0.2	0.17	0.08	-0.01		
	전체 단어	42420	11812	3391	981	280	77.4	19.8	4.2	0.39	-0.298	
이 름		41713	10042	2623	727	211	64	19.8	6.1	1.8	0.49	-0.04

$$P(Z \geq x) = 1 - P(Z < x)$$
$$= 1 - \sum_{i=1}^{x} \binom{x_0}{i} \left(\frac{1}{1+t}\right)^i \left(1 - \frac{1}{1+t}\right)^{x_0 - i}$$

위 식에서 Z는 모수가 $\frac{1}{1+t}$인 이항분포를 따른다. 즉 $Z \sim Bin\left(x_0, \frac{1}{1+t}\right)$이다. $\hat{\Delta}$의 추정을 위해 말뭉치의 단어와 인명자료의 표본을 이용하여 추정하였다. 한편 말뭉치 단어를 등록체언, 등록복합명사, 미등록체언, 등록용언, 미등록용언, 보조용언, 부사, 조사, 기타 등 9개의 그룹으로 분류하여 각 그룹의 개수를 추정한 후 말뭉치 전체단어의 추정과 비교하였다. 각 항목별 추정치의 수렴과정은 〈그림 1〉에서 볼 수 있으며(지면관계상 말뭉치의 전체 단어와 미등록체언에 대한 그림만 실었다), 추정치는 〈표 3〉에 정리되어있다.



(그림 1) 전체 단어와 미등록체언의 수렴과정

(Fig. 1) Convergence process fro corpus words and unregistered substantives

〈표 3〉 각 그룹별 단어 및 이름의 추정 결과
〈Table 3〉 Estimation result of words and names

		총출현개수	서로다른개수	x_0	$\hat{\triangle}$
말뭉치단어	등록체언	3,401,024	35,074	9	32,248
	등록복합명사	168,369	56,272	10	203,710
	미등록체언	671,098	183,091	8	525,428
	등록용언	1,666,942	6,259	9	9,567
	미등록용언	3,167	1,758	5	7,036
	보조용언	74,353	187	7	269
	부사	12,714	741	4	824
	조사	2,576,482	356	1	26
	기타	299,441	1,891	7	1411
	전체단어	8,872,590	285,629	9	776,763
이름		3,653,772	456,453	10	1,036,550

t가 증가함에 따른 추정치의 변화는 〈그림 1〉에서 보는 바와 같이 항목별로 차이가 있으나 t의 값이 작을 때는 급격히 증가하나 t의 값이 100을 초과하는 경우 약간의 진동은 있으나 극한값에 수렴함을 볼 수 있다. 즉 t가 증가함에 따라($t \rightarrow \infty$) 각 항목의 추정치는 극한값에 수렴하게되는데 이 극한값이 추정치가 된다. 말뭉치 단어의 경우, $t \rightarrow \infty$일 때 수렴값은 776,763이며 미등록체언의 경우, $t \rightarrow \infty$일 때 수렴값은 525,428이다. 각 항목별 추정치는 〈표 3〉에 정리되어 있다.

추정결과 말뭉치 전체 단어의 경우 285,629개의 현재 말뭉치에서 사용된 단어 외에 적어도 776,763개의 사용되지 않은 단어가 더 존재하는 것으로 추정되었다. 이 추정치는 말뭉치의 단어를 9개의 그룹으로 분류하여 각 그룹별로 추정한 추정치의 합인 780,510개와 근사한 결과를 보였다. 합계에서의 차이는 ξ_y의 수렴속도에 따른 추정치의 변화에 의한 것이다. 이름의 경우 주어진 456,453개의 이름 외에도 1,036,550개의 이름이 더 존재하는 것으로 추정되었다.

〈표 3〉의 추정치에 대한 분산은 각 항목에 대해 독립적인 포아송 절차를 따른다는 가정 하에 구할 수 있다. 즉 $\hat{\triangle}$ 의 분산은 다음과 같다.

$$var(\hat{\triangle}) = \sum_{x=1}^{\infty} h_x^2 \eta_x \quad \text{----------(10)}$$

식(10)을 이용한 각 항목별 표준편차(standard deviation), $\sqrt{var(\hat{\triangle})} = \sqrt{\sum_{x=1}^{\infty} h_x^2 \eta_x}$는 〈표 4〉에 정리되어있다.

〈표 4〉에서 보는 바와 같이 이름의 경우 $x_0 = 10$, $t = 170$일 때 $\hat{\triangle}^{10}(170) = 1,036,550 \pm 57,196$으로 추정치 1,036,550에 대한 표준편차가 57,196개로 표준편차에 대한 추정치의 비율이 5.5%이다. 말뭉치 단어의 경우 $x_0 = 9$, $t = 9$일 때 $\hat{\triangle}^{9}(170) = 776,763 \pm 21,242$로 추정치 776,763에 대한 표준편차가 21,242개로 표준편차에 대한 추정치의 비율이 2.7%이다. 그러나 말뭉치 단어 중에서 미등록체언의 경우 $\hat{\triangle}^{8}(170) = 525,428 \pm 9,906$으로 표준편차에 대한 추정치의 비율이 ±1.88로 정확도(precision)가 상당히 높다. 그러나 등록용언의 경우 $x_0 = 9$, $t = 170$일 때 $\hat{\triangle}^{9}(170) = 9,567 \pm 3,047$로 추정치 9,567에

〈표 4〉단어 및 이름의 표준편차
〈Table 4〉 Standard deviation of words and names

		총출현개수	서로다른개수	$\hat{\triangle}$	표준편차
말뭉치단어	등록체언	3,401,024	35,074	32,248	8,282
	등록복합명사	168,369	56,272	203,710	16,280
	미등록체언	671,098	183,091	525,428	9,906
	등록용언	1,666,942	6,259	9,567	3,047
	미등록용언	3,167	1,758	7,036	501
	보조용언	74,353	187	269	80
	부사	12,714	741	824	91
	조사	2,576,482	356	26	5
	기타	299,441	1,891	1411	1,078
	전체단어	8,872,590	285,629	776,763	21,242
이름		3,653,772	456,453	1,036,550	57,196

대한 표준편차가 3,047로 지나치게 크므로 추정의 의미가 없어진다. 같은 현상이 기타에서도 나타나는데 $x_0=7$ 일 때 $\hat{\triangle}^7(170)=1,411 \pm 1,078$로 표준편차가 추정치와 비슷하다. 이와 같은 현상은 〈표 2〉에서 볼 수 있듯이 ξ_y의 값이 '0'에 수렴하는 속도가 느린 경우에 나타남을 알 수 있다. 그러나 말뭉치 단어의 각 그룹을 합한 전체 단어의 경우 수렴속도가 빠르기 때문에 추정치의 정확도가 높다고 볼 수 있으며 각 그룹별 추정치의 합보다는 전체단어에 대한 추정치의 정확도가 높다고 볼 수 있다. 한편 오일러 변환을 이용한 일반화된 추정식은 오일러 변환식의 ξ_y 값이 '0'에 수렴하는 단계가 느린 경우는 적절치 않음을 〈표 2〉를 통해 파악하였으며 '0'에 수렴하는 속도가 느린 경우에 표준편차를 감소시키는 방법에 대한 연구가 필요하다고 하겠다.

4. 결 론

본 논문에서는 한국어의 어휘량을 예측하기 위하여 대규모 말뭉치와 대규모 인명자료를 기반으로 하여 사용되지 않은 단어나 이름의 개수를 추정하는 방법을 제시하였으며 제시된 방법을 통하여 한국어 단어의 총 개수와 한국사람 이름의 총 개수를 추정하였다. 추정된 결과를 보면 단어가 약 106만개, 이름의 개수가 149만 개이었다. 이를 통하여 한국어에서 사용되는 총 단어의 개수가 고유명사를 포함하여 약 300만개 이상임을 예상할 수가 있다. 물론 형태가 같으나 의미가 다른 단어를 별도로 추정할 경우 개수는 더욱 증가할 것이다.

말뭉치를 이용한 한국어 단어 개수 추정에 있어서의 가장 중요한 문제중의 하나는 말뭉치에서 정확한 단어를 분석하여 추출하는 일이다. 이를 위하여 품사가 부착된 대규모 말뭉치가 구축되거나 오류가 없는 말뭉치에서 정확하게 형태소를 분석할 수 있는 형태소 분석기의 개발이 필요하다. 그러나 아직 품사정보가 부착된 대규모 말뭉치는 구축되지 않은 상태이며 정확한 형태소 분석기가 개발되지 않은 상태이다. 본 논문에서 사용된 1000만 어절 말뭉치는 오류가 완전히 제거되지 못한 말뭉치였고 형태소 분석기도 구문분석이나 의미분석이 가미되지 못한 어절 단위의 형태소 분석기이므로 분석된 단어의 정확도가 100%가 되지 못하였고 이를 수정하는데 많은 노력과 시간이 소모되었다.

본 연구에서는 빈도를 이용한 일반화된 선형모형을 적용하여 한국어 단어의 총개수를 추정하였는데 오일러 변환식의 수렴속도가 느린 경우에 표준편차를 줄이는 방법이 연구되어야할 것으로 보인다. 이는 사전결합 밀도함수(joint prior density function)를 이용한 베이지안 접근법의 이용으로 해결할 수도 있을 것으로 기대되나 베이지안 접근법을 이용하기 위해서는 각 항목별 사전 확률 밀도함수의 형태를 파악해야 하는 어려움이 있다. 이외에도 포획-재포획(capture-recapture) 모형을 이용하여 총 개수를 추정하는 방법도 고려해볼 수 있을 것이다.

참 고 문 헌

[1] 강승식, "음절 정보와 복수어 단위정보를 이용한 한국어 형태소 분석", 서울대학교 공학박사 학위논문, 1993.

[2] 김광해, 국어 어휘론 개설, 집문당, 1993.

[3] 김봉섭, 이종혁, 이근배, "말뭉치를 기반으로 한 한국어 명사의 의미 중의성 해소," 한국정보과학회 가을학술발표논문지, 1997.

[4] 김충희, 정인상, 이상호, "퍼스널 컴퓨터를 이용한 국어자료 처리의 효율적 방안 연구", 한국어전산학, 창간호, 1991.

[5] 이상섭, "말뭉치, 그 개념과 구현", 사전 편찬학 연구 5·6집, 1995.

[6] 정영미, "국어 어휘의 통계적 특성과 이의 응용", 사전편찬학 연구 5·6집, 1995.

[7] 최기선, 박동인, "국어정보베이스의 현재와 미래", 한국정보과학회 정보과학회지, 제15권 10호, 1997, 10.

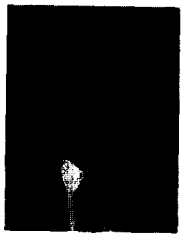
[8] Efron, B., Thisted, R., "Estimating the Number of Unseen Species : How Many Words did Shakespeare Know?" Biometrika, Vol.63, 3, pp.435-447, 1976.

[9] Efron, B., Thisted, R., "Did Shakespeare Write a Newly-Discovered Poem?" Biometrika Vol.74, No.3, pp.445-455.

[10] Good, I. J , Toulmin, G. H., "The Number of New Species, and the Increase in Population Coverage, When a Sample is Increase," Biometrika, Vol.43, pp.45-63, 1956.

[11] Fisher, R. A., "Statistical Methods for Research Workers," 14th edn. Oliver and Boydi Edinverg, 1970.

[13] McNeil, D., "Estimating an Author's Vocabulary," Journal of American. Statistical. Assoc, 68., pp.92-96, 1973.



김 성 기

1983년 서울대학교 컴퓨터공학과 졸업

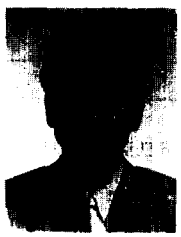
1985년 서울대학교 대학원 컴퓨터공학과(공학석사)

1992년 서울대학교 대학원 컴퓨터공학과(공학박사)

1985년~1987년 삼성전자 연구원

1993년~현재 한신대학교 전자계산학과 부교수

관심분야 : 데이터베이스, 데이터 모델링, 한국어 정보처리



한 근 식

1983년 고려대학교 통계학과 졸업

1990년 Iowa State Univ. 통계학과 졸업(석사)

1993년 Oklahoma State Univ. 통계학과 졸업(박사)

1994년~현재 한신대학교 전산통계학과 교수

관심분야 : 패턴인식, 한국어 정보처리, 통계 데이터베이스