어절 빈도 조사에 의한 최적의 고빈도 어절 집합 추출

강승식
국민대학교 컴퓨터학부

Extracting High-Frequency Optimal Korean Word Set by Word Frequency Statistics

Seung-Shik Kang
School of Computer Science, Kookmin University, Seoul, Korea
sskang@kookmin.ac.kr, http://nlp.kookmin.ac.kr/

요 약

1500만, 700만, 100만 어절 크기의 세 가지 원시 말뭉치로부터 한국어 어절 빈도를 조사하였다. 각 말뭉치에 대한 어절 빈도 결과를 비교-분석하여 활용가치가 높은 고빈도 어절 집합을 구하였다. 고빈도 어절 집합의 효용성을 검증하기 위해 일반문서에 대한 어절 적중률을 실험하였다. 그 결과로 고빈도 563 어절이 24.5%, 9484 어절이 51.5%, 184246 어절이 81.6%의 어절 적중률을 보였다.

1. 서론

한글 음절의 빈도를 조사해 보면 그 출현빈도에 따라 고빈도 음절과 저빈도 음절로 구분된다. 또한, 조사 혹은 어미에는 소수의 특정 음절들이 주로 사용되고 있다. 이러한 한글 음절 빈도 및 음절 특성은 한국어 형태소 분석기를 비롯한 한국어 정보처리 시스템을 개발하는데 활용되고 있다[1].

음절빈도와 마찬가지로 한글 어절의 출현빈도를 조사해 보면 고빈도 어절 및 저빈도 어절로 구분된다. 소규모의 고빈도 어절 집합이 주어지면 형태소 분석, 태깅, 자동 띄어쓰기 등 한국어 정보처리 시스템의 성능 및 정확도를 향상시키는데 활용될 수 있다. 김재한(1994)은 이러한 어절 빈도 특성을 이용하여 형태소 분석기의 성능을 향상시키는 연구를 수행하였는데, 이 연구에서는 350만 어절 크기의 원시 말뭉치로부터 추출하였고, 어절 적중률이 75.1%인 고빈도 어절 15만 개를 사용하였다[2].

본 논문에서는 고빈도 어절 집합이 특정 분야 문서에 제약을 받지 않도록 3개의 원시 말뭉치로부터 출현빈도를 각각 계산하여 문서의 유형과 무관한 고빈도 어절 집합을 구하였다. 고빈도 어절 집합의 객관성을 평가하기 위해 다양한 문서들에 대해 어절 적중률을 실험하여 그 활용 가치를 평가하였다.

2. 말뭉치 구성 및 어절 추출

한국어의 어절 빈도를 조사하기 위해 표 1과 같이 세 가지 유형의 원시 말뭉치(raw corpus)를 수집하였다. 말뭉치-1은 '21세기 세종계획'에서 구축된 균형 말뭉치로서 1998년 및 1999년 세종계획 프로젝트에서 구축된 2개의 말뭉치를 합한 것이다[3,4]. 말뭉치-2는 KTSET-1, KTSET-2, Krist collection, 신문기사, 백과사전, 초등학교 교과서 등으로 구성되어 있다[5]. 그리고 말뭉치-3은 자연언어 처리를 위해 구축된 100만 어절 규모의 말뭉치이다.[1)]

표 1. 말뭉치의 크기(어절수)

말뭉치 / 어절수	말뭉치-1	말뭉치-2	말뭉치-3
순수한글 어절수 (원문 어절수)	14,003,474 (15,231,638)	6,778,965 (7,185,665)	977,394 (1,030,639)
중복제거 어절수	1,494,646	926,495	171,455

어절 빈도를 조사하기 위해 원시 말뭉치로부터 순수한 한글 어절을 제외한 영문, 숫자, 한-영 혼합 어절을 제거하였다. 표 1에서 '원문 어절수'는 원시 말뭉치의 어절 개수이고, '순수 한글 어절수'는 원시 말뭉치에서 영문, 숫자, 한-영 혼합어 등을 제외하고 순수하게 한글 음절로만 구성된 어절 개수이다.

'순수 한글 어절'을 구하기 위해 괄호 안의 글자를 제거하였으며, 'abc 를'과 같이 영문자와 조사/어미에 공백이 삽입된 것은 전처리 과정에서 "아스키문자 + 공백문자 + 한글" 유형에 대해 공백 문자를 제거하였다. 그렇지 않을 경우에 조사 '은/는/이/가/을/를' 등의 빈도가 매우 높아지기 때문이다.

1) 각 원시 말뭉치에서 원문의 일부가 중복되었는지는 확인되지 않았다.

3. 어절 빈도 조사

세 가지 유형의 말뭉치에 대해 어절 빈도를 조사하고 고빈도 어절 순으로 정렬하여 누적 빈도 및 누적 어절수를 계산하였다. 각 말뭉치에 대한 어절 빈도를 조사한 결과는 표 2, 표 3, 표 4와 같다.

표 2. 말뭉치-1의 어절 빈도

빈도수 / 누적빈도	빈도수	누적 어절수
10%	14,172 이상	44
20%	4,423 이상	235
30%	1,555 이상	793
40%	620 이상	2,297
50%	270 이상	5,815
60%	113 이상	14,108
70%	43 이상	34,870
80%	13 이상	99,047
85%	7 이상	169,196
90.5%	3 이상	369,394
93.5%	2 이상	578,645
100%	1 이상	1,494,646

표 3. 말뭉치-2의 어절 빈도

빈도수 / 누적빈도	빈도수	누적 어절수
10%	7,134 이상	40
20%	2,167 이상	221
30%	745 이상	790
40%	309 이상	2,275
50%	139 이상	5,652
60%	60 이상	13,295
70%	23 이상	32,295
80%	7 이상	93,871
85.2%	4 이상	159,680
87.7%	3 이상	215,708
91.3%	2 이상	338,827
100%	1 이상	926,495

표 4. 말뭉치-3의 어절 빈도

빈도수 / 누적빈도	빈도수	누적 어절수
10%	1,411 이상	37
20%	498 이상	165
30%	206 이상	480
40%	95 이상	1,206
50%	44 이상	2,781
60%	21 이상	6,118
70%	9 이상	14,393
80%	4 이상	32,078
84.1%	3 이상	43,355
89.7%	2 이상	70,354
100%	1 이상	171,455

말뭉치-1과 말뭉치-2의 누적 어절수는 누적 빈도 10%에서 90%(빈도수 2이상)까지 그 개수가 매우 유사함을 알 수 있다. 이에 비해, 말뭉치-3은 누적 어절수의 차이가 크다. 그 이유는 말뭉치의 크기가 작기 때문일 것으로 추정된다.

말뭉치-1과 말뭉치-2로부터 추출된 고빈도 어절 집합에 대해 일치되는 어절들을 조사하였다. 표 5는 말뭉치-2의 고빈도 어절수를 기준으로 일치율을 조사한 것으로, 말뭉치-1과 말뭉치-2의 고빈도 어절을 동일한 개수씩 추출하여 공통 어절수를 조사한 것이다. 즉, 누적 빈도 20%의 경우 말뭉치-1과 말뭉치-2에서 각각 고빈도어 221개씩 추출했을 때 일치된 어절수가 162개이다.

표 5. 말뭉치-1, 말뭉치-2의 고빈도어 일치율

어절수 / 누적빈도	고빈도 어절수	일 치 어절수	일치율(%)
10%	40	29	72.5
20%	221	162	73.3
30%	790	562	71.1
40%	2,275	1,556	68.4
50%	5,562	3,911	69.2
60%	13,295	9,253	69.6
70%	32,295	22,203	68.8
80%	93,871	60,607	64.6
85%	159,680	96,771	60.6
90%	338,827	177,503	52.4

표 6은 누적 백분율 단위로 말뭉치-1의 누적 어절수에 말뭉치-2의 누적 어절수가 포함된 개수를 조사한 결과이다. 즉, 누적 빈도 20%의 경우 말뭉치-2의 누적 어절수 221개 중에서 말뭉치-1의 누적 어절수 235개에 포함된 어절수가 165개이다.

표 6. 말뭉치-1에 대한 말뭉치-2의 포함비율

어절수 / 누적빈도	고빈도 어절수	포 함 어절수	포함율(%)
10%	40	30	75.0
20%	221	165	74.7
30%	790	563	71.3
40%	2,275	1,562	68.7
50%	5,562	3,953	69.9
60%	13,295	9,484	71.3
70%	32,295	22,906	70.9
80%	93,871	62,044	66.1
85%	159,680	99,213	62.1
90%	338,827	184,128	54.3

4. 실험 및 평가

4.1 고빈도 어절 집합의 선정 실험

고빈도 어절 집합은 대용량의 균형 말뭉치로부터 조사된 빈도수에 따라 선정할 수 있다. 본 논문에서 수집한 3개의 말뭉치 중에서 말뭉치-1은 균형 말뭉치이고 어절수가 가장 많으므로 비교적 신뢰도가 높은 말뭉치이다. 따라서 문서 유형과 무관하게 적용 범위(어절 적중률)가 넓은 고빈도 어절 집합은 말뭉치-1에 대해 조사된 빈도수에 따라 구하는 것이 적합할 것으로 추정된다.

그러나 말뭉치-1의 1500만 어절 규모가 어절 적중률이 높으면서 "최소 크기의 고빈도 어절 집합"을 구하는데 적절한지는 명확하지 않다. 말뭉치-1의 고빈도 어절수(표 2)와 말뭉치-2의 고빈도 어절수(표 3)를 누적 빈도별로 비교해 보면 말뭉치의 크기가 약 2배임에도 불구하고 매우 유사하다. 또한, 말뭉치-2는 균형 말뭉치인지가 검증되지 않았고 그 크기도 말뭉치-1의 2분의 1 정도이므로 말뭉치-1에 대한 빈도 조사 결과를 기준으로 고빈도 어절 집합을 구하는 것이 최선일 것이다.

말뭉치-1의 빈도 조사의 신뢰도를 검증하고 최소 크기의 고빈도 어절 집합을 선정하기 위하여 말뭉치-1의 고빈도 어절 집합(표 2)과 말뭉치-1의 어절 집합에서 말뭉치-2와 공통인 어절 집합(표 6)에 대해 어절 적중률 실험을 하였다. 실험에 사용된 어절수는 누적 빈도 90%에 속하는 어절들이며, 실험 말뭉치의 유형 및 어절수는 표 7과 같다. 이 문서들은 말뭉치-2에 속한 문서 중에서 초등학교 교과서를 제외한 것이다.[2)]

표 7. 실험 말뭉치의 유형 및 어절 적중률

적중률 / 문서유형	어절수	어절 적중률 (말뭉치-1 369,394 어절)	어절 적중률 (공통 어절수 184,128 어절)
백과사전	1,118,094	88.4	87.7
신문칼럼1	1,763,619	88.9	87.7
신문칼럼2	891,045	86.0	84.6
문학작품	301,212	78.3	77.0
신문기사	469,077	75.7	74.8
KTSET	131,479	83.5	83.5
KTSET2	585,996	85.1	84.8
KRIST	1,277,900	72.4	72.2
합계/평균	6,538,422	82.3	81.6

표 7의 실험 결과에 의하면 어절 집합의 크기가 2배 차이인데 비해 어절 적중률은 0.7% 차이로서 말뭉치-1과 말뭉치-2에 공통되는 어절 집합이 더 효용가치가 높음을 알 수 있다.[3)]

4.2 누적 빈도별 어절 적중률

한국어 정보처리 시스템에서 활용 가치가 높은 고빈도 어절 집합은 "최소의 어절수로 최대의 어절 적중률"을 보이는 어절들로 구성되어야 한다. 표 6에서 구한 말뭉치-1과 말뭉치-2의 공통 어절들에 대해 누적 빈도별 어절 적중률을 실험하였으며 그 결과는 표 8과 같다.

표 8. 고빈도 어절 집합의 어절 적중률

어절 적중률 / 고빈도 어절수(개)	어절 적중률(%)
30	7.7 %
165	16.0 %
563	24.5 %
1,562	32.7 %
3,953	41.8 %
9,484	51.5 %
22,906	61.5 %
62,044	72.1 %
99,213	76.5 %
184,128	81.6 %

5. 결론

소수의 고빈도 어절이 문서에서 자주 출현하는 비율 및 고빈도 어절 집합을 구하기 위해 대용량 원시 말뭉치로부터 어절 빈도를 조사하였다. 다양한 문서들에 대한 어절 적중률 실험에 의해 소수의 고빈도 어절들이 한글 문서에서 자주 출현함을 확인하였다.

어절 적중률이 높으면서 어절수가 최소인 고빈도 어절 집합을 구하기 위해 두 가지 말뭉치에 공통인 고빈도 어절 집합을 구하였다. 이 어절 집합에 대한 어절 적중률 실험 결과로 고빈도 30개 어절이 어절 적중률 7.7%, 563 어절이 24.5%, 9,484 어절이 51.5%, 184,128 어절이 81.6%로 나타났다.

고빈도 어절 집합은 한국어 정보처리 시스템의 성능과 정확도를 높이는데 활용될 수 있다. 형태소 분석기의 처리 속도를 향상시키기 위하여 고빈도어 9,484개에 대

2) 초등학교 교과서는 어절 적중률이 다른 모든 문서들에 비해 최소 3% 이상 높아서 평균 적중률을 계산하는데 적합하지 않다고 판단되었기 때문이다.

3) 세 개의 원시 말뭉치에 공통된 어절 집합에 대해서도 그 효용성을 실험하였으나 90%에 대한 공통 어절수가 70,354개인데 비해 66.4%로 어절 적중률이 매우 낮은 결과를 보였다. 이는 말뭉치-3의 크기가 어절 빈도를 계산하는데 너무 작기 때문인 것으로 추정된다.

해 기분석 사전을 구축하면 51.5%의 적중률을 기대할 수 있으며, 어절수를 184,128개로 확대하면 기분석 사전의 적중률이 81.6%가 된다. 또한, 품사 태깅에서는 고빈도 어절을 중심으로 태깅 정확도를 높이는 방안을 강구할 수 있으며 맞춤법 검사, 음성인식, 문자인식 시스템에서 오류어의 발견 및 교정하는데 유용할 것으로 기대된다.

6. 참고 문헌

[1] 강승식, *음절정보와 복수어 단위정보를 이용한 한국어 형태소 분석*, 서울대학교 박사학위 논문, 1993.

[2] 김재한, 옥철영, "어절 사전을 이용한 한국어 형태소 분석", 한국정보과학회 봄 학술발표 논문집, pp.813-816, 1994.

[3] 문화관광부, "21세기 세종계획 균형 말뭉치 '98", 국립국어연구원, 1998.

[4] 문화관광부, "21세기 세종계획 균형 1999 세종 말뭉치", 국립국어연구원, 1999.

[5] 김재군, 김영환, 김성혁, "한국어 정보검색연구를 위한 시험용 데이터 모음(KTSET) 개발", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.378-385, 1994.

부록. 말뭉치-1, 2의 고빈도 어절 집합(누적 빈도 10%)

빈도수	누적빈도	어절	빈도수	누적빈도	어절	공통
101854	0.00727	그	58020	0.00856	있다	있다
94028	0.01399	수	47677	0.01559	수	수
87748	0.02025	있다	44871	0.02221	그	그
80408	0.02600	있는	42567	0.02849	있는	있는
80378	0.03174	이	34014	0.03351	이	이
58414	0.03591	것이다	31761	0.03819	및	것이다
52544	0.03966	한	30679	0.04272	것이다	한다
36371	0.04226	것은	23311	0.04616	한다	한
36161	0.04484	그러나	21845	0.04938	한	대한
34638	0.04731	것이	19893	0.05231	대한	우리
32514	0.04963	한다	17250	0.05486	우리	것이
31907	0.05191	대한	15422	0.05713	것이	때
31659	0.05417	하는	14724	0.05931	때	할
31547	0.05643	등	14538	0.06145	본	하는
30669	0.05862	할	14492	0.06359	할	것은
28330	0.06064	있었다	14217	0.06569	하는	것으로
27756	0.06262	또	14136	0.06777	것은	또
27570	0.06459	하고	13781	0.06980	것으로	그러나
27134	0.06653	같은	13112	0.07174	위한	같은
26448	0.06842	나는	12403	0.07357	또	것을
24565	0.07017	것을	11637	0.07528	그러나	등
24481	0.07192	때	11420	0.07697	따라	위해
23435	0.07359	것	11240	0.07863	같은	그리고
22561	0.07520	우리	10904	0.08024	것을	하고
21914	0.07677	그리고	10873	0.08184	등	가장
21828	0.07833	없는	10495	0.08339	위해	있었다
21695	0.07988	것으로	10336	0.08491	그리고	했다
21406	0.08141	그는	9294	0.08628	하고	없다
20713	0.08288	했다	8831	0.08759	가장	없는
19030	0.08424	더	8620	0.08886	있었다	다른
18215	0.08554	때문에	8293	0.09008	했다	
18029	0.08683	다른	8164	0.09129	또한	
17649	0.08809	두	8095	0.09248	없다	
16722	0.08929	다시	7917	0.09365	없는	
16020	0.09043	없다	7887	0.09481	많은	
16013	0.09157	아니라	7420	0.09591	다른	
15863	0.09271	이런	7381	0.09699	큰	
15800	0.09384	그의	7314	0.09807	이를	
15485	0.09494	위해	7232	0.09914	하여	
15407	0.09604	내	7133	0.10019	여러	
14878	0.09710	그런				
14418	0.09813	함께				
14398	0.09916	어떤				
14272	0.10018	가장				