어절 내 형태소 출현 정보와 클러스터링 기법을 이용한 어휘지식 자동 획득

유원희† · 서태원†† · 임희석†††

요 약

본 논문은 자연어처리 연구를 위하여 지도학습(supervised learning)방식의 어휘지식(lexical knowledge) 수동 구축 방법의 한계점을 극복하기 위하여 비지도학습(unsupervised learning)방식의 자동 어휘지식 획득 모델을 제안한다. 제안하는 모델은 벡터화, 클러스터링, 어휘지식 획득 과정을 통하여 입력으로 주어지는 어휘목록에서 어휘지식을 자동으로 획득한다. 모델의 어휘지식 획득 과정에서 파라미터 변화에 따른 어휘지식 개수의 변화와 어휘지식의 특징이 나타나는 어휘지식 사전의 일부 모습을 보인다. 실험결과 어휘지식 중 하나로 획득되는 어휘범주 지식의 클러스터가 일정한 개수에서 수렴하는 것이 관찰되어 어휘지식을 필요로 하는 전자사전 자동구축의 가능성을 확인하였다. 또한 한국어 특성이 반영되어 좌·우 통사정보가 포함된 어휘사전을 구축하였다.

주제어 : 어휘지식, 자동획득, 클러스터링

The automatic Lexical Knowledge acquisition using morpheme information and Clustering techniques

Wonhee Yu† · Taewon Suh†† · Heuiseok Lim†††

ABSTRACT

This study offered lexical knowledge acquisition model of unsupervised learning method in order to overcome limitation of lexical knowledge hand building manual of supervised learning method for research of natural language processing. The offered model obtains the lexical knowledge from the lexical entry which was given by inputting through the process of vectorization, clustering, lexical knowledge acquisition automatically. In the process of obtaining the lexical knowledge acquisition of model, some parts of lexical knowledge dictionary which changes in the number of lexical knowledge and characteristics of lexical knowledge appeared by parameter changes were shown. The experimental results show that is possibility of automatic building of Machine-readable dictionary, because observed to the number of lexical class information cluster collected constant. also building of lexical ditionary including left-morphosyntactic information and right-morphosyntactic information is reflected korean characteristic.

Keywords : Lexical Knowledge, Automatic Acquisition, Clustering

† 정 회 원: 고려대학교 컴퓨터교육과 박사과정
†† 종신회원: 고려대학교 컴퓨터교육과 교수
††† 종신회원: 고려대학교 컴퓨터교육과 교수(교신저자)
논문접수: 2010년 1월 6일, 심사완료: 2010년 1월 25일
* 이 논문은 2009년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2009-32A-H00017)

1. 서 론

지금까지 자연어처리 연구는 정보검색, 기계번역, 텍스트 데이터마이닝, 문서요약 및 대화시스템, 음성인식 등의 다양한 연구 성과를 보이고 있다. 다양한 연구 성과들은 컴퓨터가 자연어의 의미를 파악하도록 하기를 원하는 지능적인 언어처리 시스템을 개발하는 것이 공통 목적이며, 시스템 개발을 위해서는 사전에 구축된 지식인 전자사전이 반드시 필요하다.

전자사전(Machine-readable dictionary)은 자동번역, 자동 정보 처리 등을 목적으로 컴퓨터에 기억하여 두는, 단어·용어·목차·색인 등의 어휘와 어휘 지식으로 구성되어 있다.

기존 전자사전의 구축에서 어휘 지식은 사람이 수작업으로 구축하고 있던 실정이다. 이러한 수작업 구축 방법은 계속해서 새롭게 생성·변형·제거되는 언어의 특성을 새롭게 반영하기 어렵고, 방대한 실생활의 언어를 모두 반영하는 크기의 전자사전을 구축하기 위해서는 너무 높은 비용과 많은 시간이 소모된다. 또한 많은 시간의 소모를 줄이기 위해서 많은 수의 사람이 작업을 하게 되면 언어의 이해도에 따라서 일정한 형태로 구축하기 힘든 문제점이 발생한다.

이러한 문제해결을 위하여 전자사전의 어휘 지식을 자동으로 획득하는 방법이 필요하다. 어휘 지식을 자동 획득 할 수 있다면 비용의 절감, 시간의 절감, 언어 변화 반영 등의 문제점을 해결할 수 있다.

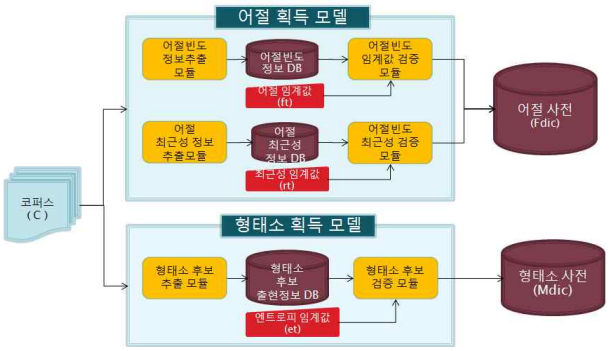
어휘 지식을 자동으로 획득하려는 해외 연구들로는 Afsaneh Fazly[1]의 문장에서 단어-의미 형태로 구성된 지식을 확률 모델을 사용하여 획득하는 연구와, Siskind[2]의 문장에서 단어-의미 형태로 구성된 지식을 룰-기반(rule-based) 모델을 사용하여 획득하는 연구, 그리고 Yu[3][4]의 인간이 학습과정에서 나타나는 여러 가지 밝혀진 효과들을 바탕으로 하여 말뭉치에서 문법을 획득하려 하였다.

본 논문에서는 기존 전자사전의 수작업 구축에서 발생하는 문제점 극복과 해외모델에서 고려할 수 없는 한국어특성을 반영한 비지도 학습 통하여 자동으로 어휘지식을 획득할 수 있는 모델을 제안한다.

2. 선행연구

기 연구에서 본 연구팀은 인간의 언어처리에서 어휘부분에 해당하는 '심성어휘집(mental lexicon)의 표상 형태에 관한 연구 모델'을 바탕으로 어휘 자동획득 시스템을 구현하였다[4]. 심성어휘집이란 인간이 언어처리를 하기 위하여 사용하는 대뇌속의 어휘집 또는 사전을 의미하고, 심성어휘집의 표상(metal lexicon representation)은 심성어휘가 대뇌 속에서 어떻게 조직화되고 저장되어 있는지를 의미한다[6][7][8][9].

심성어휘집 표상에 대한 연구는 결합 모델(full-list model), 분해모델(decomposition model), 하이브리드 모델(hybrid model)로 구분할 수 있다[10][11][12]. 본 연구팀은 하이브리드 모델을 기반으로 하는 한국어 어휘 자동획득 시스템을 구현 하였다. <그림 1>은 본 연구팀에서 개발한 한국어 자동 어휘 획득 시스템이다.



<그림 1> 한국어 자동 어휘 획득 시스템

시스템은 두 가지 모듈로 나뉘는데 하나는 어절 획득 모듈이고 하나는 형태소 획득 모듈이다. 어절 획득 모듈에서는 어절을 획득하는데, 어절 획득 원리는 빈도정보를 고려한 획득과 최근성 정보를 고려한 획득 두 가지로 구성된다.

언어생활에서 고빈도 어절은 저 빈도 어절에 비하여 인식 속도가 빠른 빈도효과(frequency effect)를 보인다[5]. 이러한 빈도 효과가 나타나는 이유 중 한 가지는 어절 전체가 분해되어 각각의 형태소가 인식되고 각 형태소를 조합한 어절이 인식되는 것이 아니라 어절 전체가 하나의 단위로 인식되기 때문이다. 이렇게 어절 전체가 사전에 저장되어야하는 이유는 어절의 인식의 효

율성에 있어서도 타당성을 갖는데, 고빈도 어절일수록 어절 전체를 사전에서 탐색하여 빠르게 인식하도록 하는 것이 언어 이해에도 효율적이라 할 수 있다.

어절 획득 모듈에서 고빈도 어절을 어절사전에 저장하여 어절 획득을 하는 것은 이러한 원리를 반영하는 것이다. 최근성 원리는 특정한 기간이나 특정한 시기에 집중적으로 반복하여 접하는 어절의 경우 결합 형태로 어절사전에 저장된다는 원리로 상대적으로 강한 자극과 입력을 받은 어절은 분해되어 이해되기 보다는 결합 형태로 이해되기 쉽기 때문이다.

어린아이의 언어 획득 과정을 살펴보면 언어생활 초기에는 어절 전체가 기억되고 이를 활용한 어절 생성을 한다. 자라면서 언어 환경에 계속 노출되고 어절 안에 반복적으로 자주 등장하는 문자열의 존재를 탐지하게 되고, 그 문자열이 어미 또는 조사 등과 같이 어울려 새로운 어절을 생성할 수 있음을 깨닫게 된다. 그 순간 그 문자열은 형태소로 심성어휘집에 등록되게 되며, 그 형태소와 결합될 수 있는 형태소에 대한 정보 및 부가 정보가 저장된다. 이러한 원리는 시스템의 형태소 획득 원리와 동일하다.

위와 같은 원리를 바탕으로 특정 문자열이 형태소가 될 수 있음을 검증하기 위하여 특정 문자열의 후행 음절의 엔트로피(sucessor entropy)와 선행 음절 엔트로피를 사용하였다. 문자열을 순방향의 음절별로 엔트로피를 측정(후행음절의 엔트로피)하여 엔트로피가 상승하는 지점을 형태소 후보로 등록하고 문자열을 역방향의 음절별로 엔트로피를 다시 측정(선행 음절 엔트로피)하여 형태소 후보를 등록한다. 양방향에서 발생한 형태소 후보가 전체 문자열과 동일하면 각각을 형태소로 보는 것이다.

어절 획득 모듈과 형태소 획득 모듈의 통합으로 한국어 자동 어휘 획득 시스템을 구성할 수 있고 시스템의 결과물로 어절 사전과 형태소사전을 얻을 수 있다. 해당 어휘 획득 시스템은 가공되지 않은 코퍼스를 입력으로 받아 어절획득 모델에서 어절을 획득하고, 형태소 획득모델에서 형태소를 획득하여 데이터베이스화하는 시스템이라고 할 수 있다. 인간의 어휘획득 과정을 모사하여 시스템을 구현하였기 때문에 인간의 초기 어휘 획득에서 나타나는 어휘폭발 현상을 관찰할 수 있었고, 결과물로 학습된 어절목록과 어휘목록을 얻을 수 있었다.

3. 어휘지식 자동 획득 모델

3.1 모델의 개요

본 논문에서 제안하는 어휘지식 획득 모델은 어절 내 형태소 출현정보와 클러스터링 기법을 통하여 입력으로 사용된 어휘목록에서 어휘지식을 획득하는 모델이다. 어휘지식이란 지식으로 사용되어지는 문자로 표현 가능한 요소들로 많은 사실, 양식, 데이터 그리고 신뢰할만한 출처로부터 일반화되어진 정보들로 구성된다. 본 논문에서 획득하는 어휘지식은 어휘범주정보, 좌측형태통사정보, 우측형태통사정보로 한정한다.

모델에서 입력으로 사용되는 형태소 사전의 데이터(D)는 획득된 어절의 집합(E), 획득된 머리형태소 집합(H), 획득된 꼬리형태소 집합(T)으로 구성된다. 획득된 머리형태소와 획득된 꼬리형태소는 획득된 어절에서 각각 좌측에 나타날 수 있는 형태소와 우측에 나타날 수 있는 형태소이다. 획득된 어절의 집합, 획득된 머리형태소 집합, 획득된 꼬리형태소 집합은 형태정보와 빈도정보로 구성된다. 따라서 형태소 사전의 데이터는 수식 1과 같이 나타낼 수 있다.

$$
\begin{aligned}
&D = [E, H, T] \\
&E = \{e_1, e_2, e_3, \cdots, e_n\} \\
&e_i = [eojeol - morph_i, frequency_i] \\
&H = \{h_1, h_2, h_3, \cdots h_n\} \\
&h_i = [head - morph_i, frequency_i] \\
&T = \{t_1, t_2, t_3, \cdots t_n\} \\
&t_i = [tail - morph_i, frequency_i]
\end{aligned}
$$

수식 1

형태소 사전 데이터의 실제 예를 살펴보면 <표 1>과 같다.

<표 1> 형태소 사전에 저장된 형태소 목록 예시

어절	어절 빈도	머리 형태소	머리 형태소 빈도	꼬리 형태소	꼬리 형태소 빈도
생각했다	1921	생각	40848	했다	100075
눈빛으로	188	눈빛	956	으로	254287
시작했다	4257	시작	18226	했다	100075
터뜨렸다	184	터뜨	688	렸다	13803
표정으로	969	표정	4452	으로	254287
지나갔다	351	지나	12113	갔다	12097
사랑으로	202	사랑	11060	으로	254287
비행기는	223	비행	2891	기는	17535
사람처럼	669	사람	74300	처럼	27541
좋아했다	230	좋아	5317	했다	100075
…		…		…	

모델의 출력결과는 선행연구로 획득된 형태소(M)에 대한 어휘지식(K)을 획득하여 구성된 어휘사전이다. 어휘지식(K)은 형태소(M)와 하나의 쌍을 이루도록 구성되며, 어휘지식은 다시 어휘범주정보(lexical class information), 좌측형태통사정보(left-morphosyntactic information), 우측형태통사정보(right-morphosyntactic information)로 구성된다.

본 논문에서 구성하는 어휘사전(Dic)은 수식 2와같이 나타낼 수 있고, 획득하는 어휘지식(k)은 수식 2와 같이 나타낼 수 있다.

$$Dic = \{[m_1, k_1], [m_2, k_2], [m_3, k_3], \cdots, [m_n, k_n]\}$$

수식 2

$$k = \begin{bmatrix} lexical\ class\ information, \\ left-morphotactic\ information, \\ right-morphotactic\ information \end{bmatrix}$$

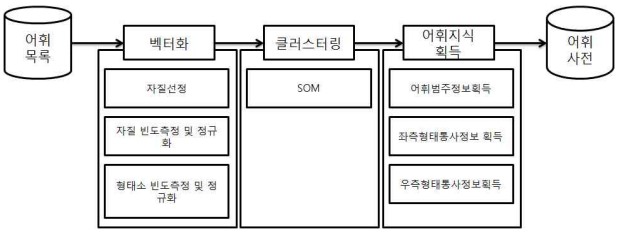
수식 3

어휘범주정보는 어휘들에서 나타나는 자질들을 바탕으로 어휘들을 클러스터링하여 나타난 클러스터로 구성된다. 좌측형태통사정보는 해당어휘의 좌측에 출현 가능한 클러스터들이고, 우측형태통사정보는 해당어휘의 우측에 출현 가능한 클러스터들로 구성된다. 예를 들어 “아버지”의 클러스터가 C_1이고, 좌측에서 나타날 수 있는 “의붓”, “수양”, “큰”, “작은”, “친”과 같은 형태통사정보의 클러스터가 L_{10}, L_{11}, L_{15}이고, 우측에서 나타날 수 있는 “를”, “가”, “께서”, “도”와 같은 형태통사정보의 클러스터가 R_3, R_7이라고 한다면 “아버지”의 어휘지식은 다음과 같이 나타낼 수 있다.

$$\text{아버지에 대한 어휘지식} = \begin{bmatrix} C_1, \\ left = \{L_{10}, L_{11}, L_{15}\}, \\ right = \{R_3, R_7\} \end{bmatrix}$$

수식 4

본 논문에서 제안하는 결과를 얻기 위하여 <그림 5>와 같은 구조를 가지는 모델을 만든다. 입력으로는 선행연구에서 획득된 어휘목록을 사용하고, 벡터화, 클러스터링, 어휘지식 획득 과정을 거쳐 어휘사전을 생성한다. 벡터화에서는 자질선정, 자질 빈도측정 및 정규화, 형태소 빈도측정 및 정규화과정을 시행하고 클러스터링에서는 SOM을 기반으로 어휘들을 클러스터링 한다. 어휘지식 획득에서는 어휘범주정보를 획득하고, 좌측형태통사정보의 획득, 우측형태통사정보의 획득한다.



<그림 2> 모델의 구조도

3.2 벡터화

입력으로 사용되는 어휘목록, 즉 형태소들은 비교 가능한 양적 데이터들이 아니다. 형태소들을 비교가능하게 표현하기 위하여 벡터화에서는 각 형태소를 특성을 가지는 n차원의 벡터형태로 표현한다. 입력데이터는 머리형태소집합(H)과 꼬리형태소집합(T)을 각각 하나의 샘플집합으로 사용한다. 벡터화를 위하여 각 형태소들을 표현할

수 있는 자질을 선정하고, 자질에 대한 빈도정보를 측정하고 측정된 빈도정보를 정규화한다. 또한 형태소빈도를 측정하고 정규화함으로써 각 형태소를 벡터화한다.

머리형태소집합(H)은 $H=\{H_1, H_2, H_3, ..., H_n\}$으로 나타낼 수 있고, 꼬리형태소집합(T)은 $T=\{T_1, T_2, T_3, ..., T_m\}$으로 나타낼 수 있다.

임의의 머리형태소는 다른 여러 꼬리형태소들과 함께 나타날 수 있고, 꼬리형태소 또한 다른 여러 머리형태소들과 함께 나타날 수 있다. 함께 나타날 수 있는 가능성이 있는 모든 형태소들을 본 논문에서는 자질로 선택하여 사용한다. 앞으로 머리형태소의 자질들은 머리형태소와 함께 나타날 수 있는 꼬리형태소의 집합(PT)으로 나타내고, 꼬리형태소의 자질들은 꼬리형태소와 함께 나타날 수 있는 머리형태소의 집합(PH)으로 나타낸다. 각 형태소는 대응되는 자질집합을 양적 측정을 통하여 n차원의 벡터로 나타낼 수 있다.

각 형태소들과 함께 나타날 수 있는 형태소들을 집합 기호로 표현하면 수식 5와같이 나타낼 수 있다.

$$H_i=\{PT_1=pt_{i1}, PT_2=pt_{i2,} PT_3=pt_{i3,}..., PT_{|PT|}=pt_{i|PT|}\}$$
$$T_i=\{PH_1=ph_{i1}, PH_2=ph_{i2}, PH_3=ph_{i3}, ..., PH_{|PH|}=ph_{i|PH|}\}$$

수식 5

H_i는 i번째 머리형태소이다. H_i는 |PT|개의 차원의 가중치 값(pt_{ij})으로 표현할 수 있다. T_i는 i번째 꼬리형태소이다. T_i는 |PH|개의 차원의 가중치 값(ph_{ij})으로 표현할 수 있다.

가중치 값(pt_{ij},ph_{ij})은 자질의 빈도를 측정하고 정규화한 값과 형태소빈도측정 및 정규화 값의 곱으로 나타낼 수 있다. 머리형태소의 가중치 값을 구하는 방법을 살펴보면, 머리형태소 H_i에서의 자질 PT_i가 출현한 빈도수를 ptf_{ij}라 하면, 자질 PT_i의 정규화빈도는 수식 6과 같이 나타낼 수 있다.

$$fh_{ij}=\frac{ptf_{ij}}{\max_l ptf_{lj}} \qquad \text{수식 6}$$

여기서 최대 값 max는 형태소 H_i에서 출현 가능한 모든 자질 중에서 가장 빈도수가 큰 자질이 되며, 자질 PT_i가 한 번도 출현하지 않았다면 fh_{ij} = 0 이다. 또한 머리형태소의 총 개수를 N이라 하고 자질 PT_i가 출현한 머리형태소 수를 n_i라고 할 때, 자질 PT_i의 형태소 빈도수는 수식 7과 같이 나타낼 수 있다.

$$mfh_i=\log\frac{N}{n_i} \qquad \text{수식 7}$$

수식 6과 수식 7은 정보검색에서 가장 일반적으로 사용하는 가중치할당 방법인 tf-idf의 형태와 유사한 형태이다. fh_{ij}와 mfh_i의 곱으로 pt_{ij}를 수식 8과같이 나타낼 수 있다.

$$pt_{ij}=fh_{ij}*mfh_i \qquad \text{수식 8}$$

꼬리형태소 T_i도 마찬가지 방법으로 가중치를 구할 수 있다. 자질 PH_i가 출현한 빈도수를 phf_{ij}라 하면, 자질 PH_i의 정규화빈도는 수식 9와 같이 나타낼 수 있고, 꼬리형태소의 총 개수를 M이라 하고 자질 PH_i가 출현한 꼬리형태소 수를 m_i라고 할 때, 자질 PH_i의 형태소 빈도수는 수식 10과 같이 나타낼 수 있다. ph_{ij}는 수식 11와같이 나타낼 수 있다.

$$ft_{ij}=\frac{phf_{ij}}{\max_l phf_{lj}} \qquad \text{수식 9}$$

$$mft_i=\log\frac{M}{m_i} \qquad \text{수식 10}$$

$$ph_{ij}=ft_{ij}*mft_i \qquad \text{수식 11}$$

생성된 머리형태소집합의 벡터는 <표 2>와 같이 나타낼 수 있다.

<표 2> 머리형태소에 대한 벡터화 예시

꼬리형태소 / 머리형태소	겠지만	겠으나	는	…	습니다
가능하	1.51	0.29	0	…	0
가다듬	0.26	1.74	0	…	0
가정부	0	0	0.18	…	0
가르시아	0	0	0.08	…	0
가버렸	0	0	0	…	0

3.3 클러스터링(clustering)

클러스터링은 벡터화한 입력데이터들을 비교하여 비슷한 특성을 보이는 입력데이터를 하나의 클러스터로 분류하는 작업이다. 어휘 수준에서 비슷한 특성을 보이는 어휘들은 문법적으로 사용되는 형태가 비슷하다. 이를 이용하여 언어의 구조적 특성으로 사용되는 어휘범주와 형태통사정보를 획득할 수 있다. 어휘의 클러스터에서 통사론적 기준에서 분류정보로 쓰이는 하나의 분류정보와 다른 어휘들 간의 결합에서 나타나는 특징들을 추출할 수 있고, 형태론적 기준에서 어휘범주를 결정할 수 있다.

클러스터링에서 사용하는 입력 데이터가 벡터화 되면 기계학습 방식의 클러스터링기법을 사용한다. 본 논문에서 사용한 클러스터링 기법은 SOM(자기조직화지도)이다. SOM은 비지도 학습 기법을 사용하는 클러스터링 기법으로 입력으로 주어진 데이터를 연관성 있는 데이터끼리 클러스터를 구성한다.

SOM의 파라미터로 경쟁 층의 그리드 맵을 3*3, 4*4, 5*5, 6*6, 7*7, 8*8형태로 변화시킨 실행 결과를 살펴보고 어휘목록에 최적화된 그리드 맵의 형태를 결정한다. 최대 반복횟수는 100000회로 하였고 학습 상수는 0.02를 사용한다. 유사도 계산의 척도는 유클리디안 거리(Euclidean distance)방법을 사용한다.

3.4 어휘지식획득

어휘지식 획득은 어휘 클러스터에서 어휘 지식인 어휘범주와 좌측형태통사정보, 우측형태통사정보를 획득하는 과정이다. 획득된 어휘지식은 어휘의 문법 부류를 나타낼 수 있고 구조적 분석에 쓰이는 사전의 형태로 구성될 수 있다.

어휘지식은 어휘범주정보와 좌측형태통사정보, 우측형태통사정보로 구성된다. 어휘의 어휘범주정보는 어휘가 속해있는 클러스터를 바탕으로 획득한다. 좌측과 우측 형태통사정보는 해당 어휘의 클러스터에서 나타날 수 있는 형태소들이 속해있는 클러스터를 선택하여 구성한다. 따라서 획득된 어휘지식은 해당어휘가 나타날 수 있는 범주정보와 해당어휘의 좌측에서 나타날 수 있는 형태소의 클러스터들, 그리고 해당어휘의 우측에서 나타날 수 있는 형태소의 클러스터들을 이진(binary)형태로 비교를 할 수 있다.

머리형태소(H_i) “아버지”에 대한 어휘지식표현의 조건이 <표 3>과 같다면, 어휘지식은 <표 4>와 같이 생성된다.

<표 3> 어휘지식 표현 조건

1. 머리형태소(H_i) 는 “아버지”
2. “아버지”와 함께 나타날 수 있는 꼬리형태소는 “는”, “가”, “께서”, “를”
3. 클러스터 구성
- C_i = {아버지, 어머니, 아저씨}
- C_j = {은, 는, 이, 가}
- C_k = {께서, 께서도, 도}
- C_l = {을, 를}

<표 4> 어휘지식의 표현

M_i = [C_i, left{null}, right{C_j,C_k,C_l}]

<표 4>에서 M_i는 임의의 어휘를 나타내고, C_i는 어휘 범주정보를 나타낸다. left{}는 M_i에서 나타날 수 있는 좌측 형태통사정보, right{}는 M_i에서 나타날 수 있는 우측 통사 정보이다.

4. 실 험

4.1 입력데이터

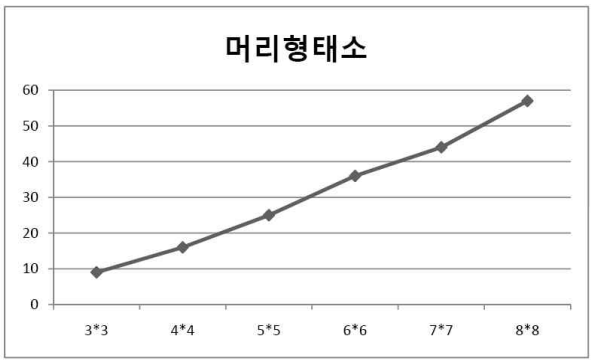
본 논문에서는 기 연구에서 획득된 어휘목록을 실험에 사용하였다. 5987개의 머리형태소와 꼬리형태소로 구성된 튜플 중에 유일하게(unique) 획득된 어휘목록인 머리형태소 3183개와 꼬리형태소 157개를 샘플 데이터로 사용하였다.

4.2 어휘목록의 벡터화

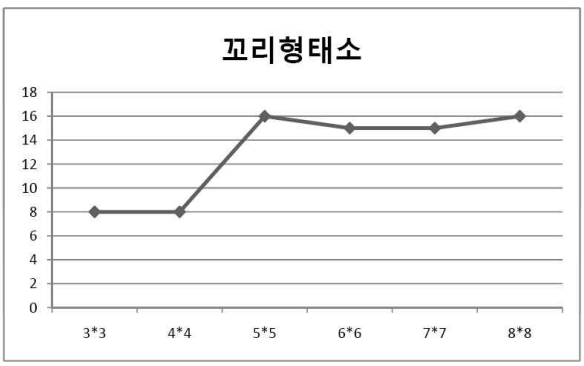
각 형태소들의 특성을 양적으로 나타내기 위하여 로우코퍼스(raw-corpus)에서 각 형태소들과 가능한 모든 조합(every possible combination)중에 해당 형태소를 잘 표현해줄 수 있는 고빈도 50개의 조합을 추출하여 형태소 벡터를 구성할 수 있는 특질(feature)로 사용하였다. 머리형태소는 40615개의 특질집합을 만들 수 있었고, 꼬리형태소는 4602개의 특질집합을 만들 수 있었다. 이러한 특질 집합을 바탕으로 머리형태소는 3183 * 40615 형태의 입력벡터를 생성하였고, 꼬리형태소는 157 * 4602 형태의 입력벡터를 생성하였다.

4.3 SOM을 사용한 클러스터링

클러스터링에 사용된 자기조직화지도 알고리즘은 그리드 맵을 3*3, 4*4, 5*5, 6*6, 7*7, 8*8로 설정하여 각 벡터 수마다 클러스터의 개수를 측정하였다. 최대 반복횟수는 100000회로 하였고 학습율은 0.02로 하였다. 유사도 계산의 척도는 유클리디안 거리(Euclidean distance)방법을 사용하였다. 가중치 벡터의 변화에 따라 머리형태소는 증가하는 클러스터 개수를 보였고 꼬리형태소는 15~16개의 클러스터 개수에서 수렴하는 형태로 클러스터링이 진행되었다. <그림 3>과 <그림 4>는 가중치 벡터 변화에 따른 머리형태소와 꼬리형태소의 클러스터 개수 변화를 보여준다. 가로축은 가중치 벡터 이고 세로축은 클러스터 개수를 나타낸다.



<그림 3> 그리드 맵 변화에 따른 머리형태소 클러스터 개수 변화



<그림 4> 그리드 맵 변화에 따른 꼬리형태소 클러스터 개수 변화

<그림 4>에서 꼬리형태소는 그리드 맵의 크기가 5*5부터 클러스터의 개수가 증가하지 않는다. 이것은 꼬리형태소는 25가지의 클러스터로 구성되는 것이 적당하다는 뜻이므로 SOM의 그리드 맵 크기를 5*5의 크기로 실험한 데이터를 사용한다. 머리형태소의 경우 그리드 맵에 상관없이 클러스터 개수가 계속 증가 하므로 꼬리형태소와 같은 그리드 개수를 선택하여 사용한다.

4.4 어휘지식 획득을 통한 사전 구성

<표 5>은 최종적으로 완성된 어휘사전의 일부 모습이다.

범주정보, 좌측통사정보, 우측통사정보의 값은 클러스터들을 나타낸 것으로 “1”로 시작하는 클러스터들은 머리형태소에서 생성된 클러스터이고, “2”로 시작하는 클러스터들은 꼬리형태소에서 생성되는 클러스터이다. 제일 앞 숫자를 제외한

나머지 숫자들은 형태소의 클러스터 정보를 나타낸다.

<표 5> 생성된 어휘사전의 일부모습

어휘	범 주 정보	좌측통사 정보	우측통사 정보
가갔	10017		20005
가격	10004		20000 20009 20012 20013
가까	10016		20000 20014
갔다	20000	10004 10006 10011 …	
…	…	…	…
나라	10006 20000	10002 10006 10008 …	20000 20005 20007 …
…	…	…	…
왔다	10016 20000	10000 10004 10006 …	20000 20003 20005 …
…	…	…	…
학교	10012 20007	10003 10004 10007 …	20000 20001 20003 …
…	…	…	…
흘러나	10017		20000
흩어	10017		20000
힘들	10010		20000 20005 20006

“가갔”, “가격”, “가까”, “흘러나“, ”흩어“, ”힘들“들의 어휘는 머리형태소에서 나타나는 어휘들이고, 각각 우측통사정보를 가진다. “나라”, “왔다”, “학교”들은 머리형태소와 꼬리형태소에서 모두 나타나는 어휘들로 각각 좌측통사정보와 우측통사정보를 모두 가진다. “갔다”는 꼬리형태소에서 나타나는 어휘이고 좌측통사정보를 가진다.

5. 결 론

본 논문에서는 한국어 어휘자동획득 시스템에서 발생하는 형태소 사전을 바탕으로 형태소의 어휘지식 자동획득을 목표로 하는 모델을 제안하였다. 어휘목록을 벡터화, 클러스터링, 어휘지식 획득 과정을 통하여 어휘지식이 자동으로 획득될 수 있는 방법을 제안하였고, 제안한 방법에 따라 실험하였다. 실험은 3183개의 머리형태소와 157개의 꼬리형태소를 입력으로 사용하여 어휘의 범주정보 및 좌측통사정보와 우측통사정보를 획득하여 사전을 구성하였다.

본 논문의 의의는 다음과 같다.

첫째, 어휘지식의 자동 획득을 통하여 전자사전의 자동구축의 가능성을 보였다. 실험에서 꼬리 형태소 부분은 어휘범주와 형태통사정보를 획득하는데 필요한 클러스터링 결과가 일정한 클러스터 개수에서 수렴하는 것이 관찰 되었다. 이것은 꼬리형태소가 일정한 도메인 크기를 가지고 있다는 것을 말해주고, 클러스터링 기법을 통하여 분류를 할 수 있다는 것을 말한다. 머리형태소가 일정한 클러스터 개수에서 수렴하지 않는 것은 차후 연구에서 고려할 것이다.

둘째, 한국어의 특성이 반영된 어휘사전을 구축하였다. 대표적인 외국어인 영어는 문장이 단어들로 구성이 되어있는데 반하여, 한국어는 머리형태소와 꼬리형태소로 문장이 구성된다. 이런 한국어의 특성상 영어와는 다른 형태의 어휘지식이 필요한데 본 논문에서는 머리형태소와 꼬리형태소의 범주정보와 좌·우측통사정보를 획득하여 한국어 분석에 필요한 어휘사전을 생성하였다.

본 연구팀은 향후 어휘지식 자동획득 모델을 발전시켜 시스템화하고 어휘자동획득 시스템과 어휘지식 자동획득을 통합하여 한국어 어휘 분석을 할 수 있는 통합 시스템 구축연구와 다른 어휘지식 정보들을 획득하여 현 시스템을 발전시킬 수 있는 연구를 지속할 계획이다.

참 고 문 헌

[1] Fazly, A., Alishahi, A., &Stevenson, A. (2007). *A probabilistic computational*

model of cross-situational word learning. To appear in Cognitive Science.

[2] Siskind, J. M. (1996). *A computational study of crosssituational techniques for learning word-to-meaning mappings.*Cognition. *61*, 39-91.

[3] Yu, C. (2005). *The emergence of links between lexical acquisition and object categorization: A computational study.* Connection Science, *17*, 381-397.

[4] Yu, C. (2006). Learning syntax--semantics mappings to bootstrap word learning. In Proc. of CogSci'06.

[5] 임희석 (2007). 한국어 어휘자동획득 시스템. **한국산학기술학회논문지. 8**, (5), 1087-1091.

[6] Bradley, D. (1980). *Lexical representation of derivational relation*, In M. Aronoff & M. L. Kean (Eds.). Cambridge. MA: MIT Press. Juncture. 37-55.

[7] Caramazza. A., Laudanna, A., &Romani, C. (1988). *Lexical access and inflectional morphology*. Cognition. *20*, 207-332.

[8] Foster, K. I. (1976). *Accessing the mental lexicon*. In R. J. Wales, E. Walker (Eds.). New approches to languge mechanisms, 257-287.

[9] Taft, M. (1991). *Reading and the mental Lexicon*. Hillsdale. NJ : Eribaum.

[10] Jung, J., Lim, H., &Nam, K. (2003). *Morphological Representations of Korean compound Nouns.* Journal of Speech and Hearing Disorders. *12,* 77-95.

[11] Lim, H., Nam, K., &Hwang, Y. (2005). *A Computational Model of Korean Mental Lexicon.* Lecture Notes in Computer Science. *3480,* 1129-1136.

[12] Nam, K., Seo, K., &Choi, K. (1997). *The word length effect on Hangul word recognition.* Korean Journal of Experimental and Cognitive Psychology. *9*, 1-18.

유 원 희

2006 한신대학교 소프트웨어학과(학사)
2009 한신대학교 컴퓨터정보학과(석사)
2009~현재 고려대학교 컴퓨터교육학과 박사과정

관심분야: 자연어처리, 인지과학, 컴퓨터 교육
E-Mail: gala@korea.ac.kr

서 태 원

1992 고려대하교 전기공학과(학사)
1995 서울대학교 전자공학과(석사)
1997 Georgia institute of technology (박사)
2008~현재 고려대학교 교수
관심분야: 임베디드시스템, 컴퓨터구조, 컴퓨터교육
E-Mail: suhtw@korea.ac.kr

임 희 석

1992 고려대학교 컴퓨터학과(학사)
1994 고려대학교 컴퓨터학과(석사)
1997 고려대학교 컴퓨터학과 (박사)
2008~현재 고려대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 자연어처리, 인지신경과학
E-Mail: limhseok@korea.ac.kr