클러스터 분석을 통한 한국어 양태부사 어순에 관한 연구

이신원*O, 황호전**, 김법균**, 안동언**, 정성종**, 두길수***
*전북과학대학 컴퓨터정보계열, **전북대학교 전자정보공학부, ***서남대학교 전기전자공학부
swlee@mail.chongin.ac.kr, {hjhwang, kyun}@duan.chonbuk.ac.kr,
{duan, sjchung}@moak.chonbuk.ac.kr, dgs@tiger.seonam.ac.kr

A Study of Korean State Adverb ordering Using Clusters

*Shin-Won Lee O, **Ho-Jeon Hwang, **Beob-Kyun Kim, **Dong-Un An, **Sung-Jong Chung, ***Gil-Su Doo
*Department of Computer Information, JeonBuk Science College,
**Department of Computer Engineering, ChonBuk National University,
***Faculty of Electric and Electronic Engineering, Seonam University

요약

본 연구에서는 영한 기계 번역 시스템의 생성단계에서 자연스러운 어순의 연속된 부사를 생성하기 위하여 클러스터링 기법을 이용하여 부사의 어순을 결정해 보고자 한다. 먼저 국문학자가 분류해 놓은 부사의 자질 정보를 살펴보고 그 자질 정보에 대한 부사의 어순을 살펴본다. 그 중에서 양태부사에 대한 어순 정보가 부사 어순 결정에 중요한 요인이 됨으로 양태 부사에 대해서만 어순을 다루기로 한다. 통합 국어정보베이스에 수록된 한국어 구문구조 부착 말뭉치를 사용하여 연속 부사를 추출하고 그 빈도수를 추출하여 부사의 자질 정보를 부여한다. 부여된 부사의 자질 정보를 가지고 부사-부사 유사도를 계산하고 이 유사도에 기반하여 양태부사들을 재분류한다. 그리고, 양태부사의 어순 비율과 클러스터링을 통해서 세분류한 어순의 비율을 제시한다.

1. 서론

영한기계번역을 위한 한국어 생성기는 하나의 번역단위마다 크게 한국어 구문구조 생성단계와 한국어형태소 생성단계로 나누어 처리한다.

생성기는 영어의 의존 구조를 입력으로 받아 한국어 어순에 맞도록 구문구조를 생성한다. 이때 영어 문장을 한국어의 어순에 맞게 생성하기 위하여, 구문 요소 ordering 기법을 사용한다. [9]

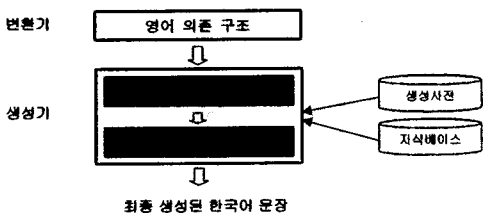

그림 1 ALKOL 시스템에서 한국어 구문구조 생성기

이러한 과정의 결과를 바탕으로 한국어 구문구조 생성기는 한국어 구문구조를 생성하고, 형태소 생성을 수행하여 최종적으로 한국어 문장을 생성한다.

다른 품사에 비하여 수식어인 부사는 기계번역 시스템에서 규칙이 없이 번역을 해 주었는데 국문학자들이 부사에 대해서 연구한 자료를 보면 부사의 자질들에 따라서 규칙이 있다는 것을 알 수 있다. 수식어인 부사가 연속해서 나올 경우 어떤 자질의 부사가 먼저 나오는지에 따라서 비문이 될 수 있고 다른 뜻을 가질 수 있음을 알 수 있다.[3]

국문학자들이 분류해 놓은 부사의 자질 정보를 가지고 통합 국어정보베이스에 수록된 한국어 구문구조 부착 말뭉치를 사용하여 부사 정보를 추출하고 그 빈도수를 추출하여 분류해 놓은 기준에 맞춰서 부사의 자질 정보를 부여해 보니 상호 연결 부사의 빈도가 많이 나옴을 알 수 있었다. 그중에서도 양태부사의 상호 연결 빈도가 많이 나옴을 알 수 있었다.

이런 문제를 해결하기 위해서 본 논문에서는 상호 연결 부사들 간의 연관 정도를 분석하기 위해서 클러스터링 알고리즘을 이용했다.

2장에서는 국문학자들이 분류한 부사의 자질 정보와 클러스터링 기법에 대해서 알아보고, 3장에서는 실험을 통해서 세분류된 양태 부사의 어순에 대해서 알아본다.

2. 부사의 자질 및 클러스터 분석

2.1 부사의 자질

부사는 주로 용언 앞에 놓여서 뒤따르는 용언을 꾸밈으로써 그 의미를 더욱 분명히 해 주는 말로서, '무엇이 어떻게 어찌한다(어떠하다)'에서 '어떻게'에 해당하는 말이다.

국문학에서 부사에 관한 연구를 보면 여러 방향으로 이루어져 왔다[1][2][3][4]. 이러한 연구들을 보면 한 문장 내에 부사가 여러 번 나올 경우에는 어순이 있다는 것을 알 수가 있다. 김창호[3]의 부사 자질을 보면 양태부사, 정도부사, 수량부사, 장소부사, 시간부사, 서법부사로 분류하였다. 여기서 양태부사는 김민수의 상태부사와 같은 자질의 부사이다.

이런 부사들의 자질 중에서 양태부사가 상호 연결되는 경

우가 많아서 본 논문에서는 양태부사의 자질만을 가지고 어순을 결정해 보기로 한다.
김창호의 각 부사의 자질을 보면 다음과 같다.

서법→화자의 태도를 나타내는 부사
시간→서술의 시점이나 시간단위를 나타내는 부사
장소→사건이 일어나는 장소나 방향을 나타내는 부사
수량→행동이나 개체의 수량을 나타내는 부사
정도→상태나 동작의 정도를 나타내는 부사
양태→행동이나 모양의 양태를 나타내는 부사

그림 2 부사의 자질

김창호는 예를 들어서 설명하고 있다. 부사의 예를 보면 다음과 같다.

(1-1) 밥을 빨리 못 먹는다.
(1-2)* 밥을 못 빨리 먹는다.
(1-3) 밥을 잘 못 먹는다.

(1-1)은 사용 가능한 문장이고 자연스럽다. (1-2)는 비문법적인 문장이다. 위의 예는 양태부사 양태부사가 연속해서 나올 경우, 부사 어순이 존재한다는 것을 알 수 있다. 연속해서 나오는 양태부사의 어순 정보를 알아보기 위해서 클러스터링 알고리즘을 이용하였다.

2.2 클러스터링 방법

사용한 클러스터링 알고리즘은 다음과 같다.

1. 임의의 클러스터를 선택한다.
2. RNN쌍이 발견될 때까지 선택한 클러스터로부터 NN체인을 따라간다. NN은 Nearest Neighbors이다.
   NN(i)=j; NN(j)=k; .. ; NN(p)=q; NN(q)=p
3. 두 클러스터를 결합하여 하나의 클러스터로 만들 때 클러스터 중심 벡터 생성한다.

$$C_{ij} = \frac{m_i \cdot C_i + m_j \cdot C_j}{m_i + m_j}$$

- *mi*는 클러스터 *Ci*에 포함된 문서의 개수
- *mj*는 클러스터 *Cj*에 포함된 문서의 개수

4. 새로 생성된 클러스터와 다른 클러스터와의 유사도를 재계산한다.

$$d_{C_k C_{i+j}} = \alpha_i \cdot d_{C_i C_k} + \alpha_j \cdot d_{C_j C_k} + \beta \cdot d_{C_i C_j} + \gamma \cdot |d_{C_i C_k} - d_{C_j C_k}|$$

$$\alpha_i = \frac{m_i + m_k}{m_i + m_j + m_k} \quad \beta = \frac{m_k}{m_i + m_j + m_k} \quad \gamma = 0$$

- *mj*는 클러스터 *Cj*에 포함된 문서의 개수

5. NN 리스트 생성한다. 오직 한 개의 클러스터가 남을 때 정지한다.
   if (NN(p)=q; NN(q)=p) exist then goto 2
   else goto 1

클러스터 분석은 N개의 항목들의 데이터 집합을 M개의 클러스터로 나누는 비계층적 방법과 항목들이나 또는 클러스터들의 쌍이 성공적으로 연결된 근접한 데이터 집합을 생성하는 계층적 방법이 있다. 그 중에서 클러스터 분석에 관한 많은 연구들이 계층적 클러스터링 방법을 이용한다.

계층적 클러스터링은 각 문서들을 각 하나의 클러스터로 하여 시작하고 유사도가 높은 두 개의 클러스터를 하나의 클러스터로 만들어 가는 과정을 반복하여 하나의 클러스터가 남을 때까지 반복한다.

Ward기법은 일반적인 계층적 클러스터링 기법을 따르고 있는데, 두 클러스터가 결합될 때 클러스터 중심에 대한 거리계산에서 전체 그룹내의 분산이 유지되는 두 클러스터를 각 단계에서 결합한다. 이때 클러스터간의 관련도는 비유사도 값을 기준으로 비유사도 값이 작을수록 두 클러스터의 관련도가 높은 것이 된다. 그래서 최소분산기법이라고도 한다.

계층적 클러스터링 방법은 N개의 문서에 대해서 2N-1개의 클러스터를 형성하고, 각 클러스터에 대한 클러스터 대표를 생성한다. 상호 최근접 이웃 (RNN : Reciprocal Nearest Neighbors) 알고리즘을 이용해서 문서를 클러스터링한다.

2.3 양태 부사 클러스터 구축

부사 클러스터를 구축하기 위해서 먼저, 각 부사를 벡터형태로 표현한 후, 부사들 사이의 유사도를 계산한다. 이 유사도에 따라 부사들을 클러스터링한다. 예를 들면 연속 부사와 빈도수 쌍을 보면 다음과 같다.

<부사 부사 빈도수> --> <아주 오래 2>

'아주' 부사와 '오래' 부사의 유사도 값을 가지고 벡터로 표현한다. 그래서, 각 부사는 연속해서 나오는 부사와 그 부사가 가지는 가중치의 쌍으로 이루어진 벡터로 표현된다. 부사의 표현을 위한 과정은 다음과 같다.

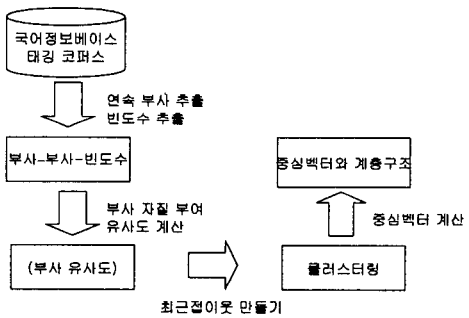

그림 3 시스템 구성도

클러스터는 여러 가지 특성을 가진 부사들로 구성되고 부사들의 부합 정도가 높을수록 두 부사는 보다 유사하다고 할 수 있다. 부사 클러스터링을 하면 유사도가 높은 부사들을 우선적으로 하나의 클러스터를 생성한다.

하나의 클러스터에 속해있는 부사들은 클러스터 중심을 형성하는 과정에서 자신들의 특성을 반영하게 된다. 즉, 클러스터의 중심이 가지는 부사와 부사의 가중치는 클러스터에 속해있는 문서들이 가지는 부사와 부사의 가중치에 의해 조정되는 것이다. 두 부사 Di와 Dj가 하나의 클러스터를 형성할 때, 클

러스터 중심을 계산하는 과정은 다음과 같다.

$$\frac{miDi + mjDj}{mi + mj}$$

여기서 m은 클러스터의 크기로 클러스터에 포함되어 있는 부사의 수를 나타낸다.

각 문서는 단어와 그 단어가 가지는 가중치의 쌍으로 표현한다. 문서 표현을 위한 과정은 통합 국어정보베이스에서 연속 부사를 추출하여 단어 빈도수를 계산하고 각 단어에 가중치를 계산한다.

3. 실험 및 평가

본 논문에서는 한국어 말뭉치를 통하여 부사의 어순을 추출하고자 1997년에 제작된 '통합 국어정보베이스'에 수록된 '한국어 구문구조 부착말뭉치' 1만여 문장을 가지고 분석하였다. 코퍼스에 나타난 부사의 자질은 2530개이고 연속 부사는 10018개이다. 세 개 연속된 부사는 140개이고 네 개 연속된 부사는 13개이다. 이 중에서 빈도수가 5번 이상 연속해서 나온 부사에 대해서만 추출하여 분석해 보았다. 이중에서 양태부사 양태부사의 빈도수는 1230이다. 이를 클러스터링 기법을 이용하여 세분류하였다. 클러스터들간의 우선순위는 앞\뒤의 관계, 뒤\앞의 관계에 따라 만들어진 클러스터의 벡터에서 값을 어느 정도 갖고 있느냐에 따라 순서를 부여할 수 있다.

84개의 양태부사를 가지고 클러스터링을 한 결과 9개의 그룹으로 분류되어 아래와 같은 결과를 보여주고 있다.

	양태1	양태2	양태3	양태4	양태5	양태6	양태7	양태8	양태9
양태1		18		18	5		20	8	8
양태2			5		8	10	26	8	13
양태3				5	6	9	10	33	6
양태4				5	46	93		21	17
양태5		18	8	14	10	22	34	45	720
양태6							16	13	75
양태7		9	5		5	23	12	8	15
양태8								11	116
양태9									13

표 1 클러스터링 결과에 따른 빈도수 추출

	양태1	양태2	양태3	양태4	양태5	양태6	양태7	양태8	양태9
양태1	0	23.4	0	23.4	6.5	0	26	10.4	10.4
양태2	0	0	7.1	0	11.4	14.3	37.1	11.4	18.6
양태3	0	0	0	7.2	8.7	13	14.5	47.8	8.7
양태4	0	0	0	2.7	25.3	51.1	0	11.5	9.3
양태5	0	2.1	0.9	1.6	1.1	2.5	3.9	5.2	82.7
양태6	0	0	0	0	0	0	15.4	12.5	72.1
양태7	0	11.7	6.5	0	6.5	29.9	15.6	10.4	19.5
양태8	0	0	0	0	0	0	0	8.7	91.3
양태9	0	0	0	0	0	0	0	0	100

표 2 클러스터링 결과에 따른 비율

양태부사에 대해서만 클러스터링을 한 결과 다음과 같은 어순을 가질 수 있다.

양태1→양태2→양태3→양태4→양태5→양태6→양태7→양태8→양태9

다음과 같은 순서로 부사의 어순을 결정할 때 보다 더 자연스러운 문장을 생성해 낼 것이다.

4. 결 론

본 논문은 영한 기계 번역 시스템에서 보다 더 자연스러운 번역을 위해서 생성할 때 국문학자가 분류해 놓은 분류기준을 토대로 '통합 국어정보 베이스'에 등록되어 있는 1만여 문장의 corpus를 분석하여 양태 부사의 자질 정보와 연속 부사를 추출하여 양태 부사의 어순 결정을 하였다.

위의 실험 결과를 보면 알 수 있듯이 양태부사를 클러스터링 기법을 이용하여 세 분류하니까 상호 연결되는 부사가 적어짐을 알 수 있었다. 이처럼 수량부사, 정도부사도 좀 더 세분화되어 분류를 하면 더 나은 부사의 어순을 생성해 낼 것으로 생각된다.

참고문헌

[1] 김민수, 국어문법론, 일조각, 1984

[2] 남기심, 표준국어문법론, 탑출판사, 1985

[3] 김창호, 국어 부사어의 어순에 관한 연구, 계명대학교

[4] 손남익, 국어 부사 연구, 박이정, 1995

[5] 박성재, 영한 번역기에서의 부사구 처리에 관한 연구, 서울대학교 컴퓨터공학과 석사학위논문, 1992

[6] 조준모외 1인, 한·영 기계 번역을 위한 부사의 위치 및 순서제약 해결의 방안 및 구현, 제 6 회 한글 및 한국어 정보처리학회, pp. 163-167, 1994

[7] 서정수, 국어문법, 한양대학교 출판원, 1996.

[8] 통합 국어정보베이스, 과학기술처, 1997.

[9] 서진원, 영한 기계 번역 시스템에서 계층적 한국어 어순 생성, 전북대학교 컴퓨터공학과 석사학위논문, 2001.

[10] http://my.netian.com/~beedman/adverd.htm

[11] J.S. Chang and K.Y. Su, "A Corpus-Based Statistics-Oriented Transfer and Generation Model for Machine Translation," TMI-93: The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, pp.3-14, 1993.

[12] S.J. Kim and N.H. Cho, "The Progress and Prospect of the 21st Century Sejong Project," Proceedings of ICCPOL 2001, Seoul, Korea, 2001.

[13] H.G. Kim and B.M. Kang, "21st Century Sejong Project - Compiling Korean Corpora," Proceedings of ICCPOL 2001, Seoul, Korea, 2001.

[14] Shin Won Lee, Dong Un An, Seong Jong Chuung, "Korean Adverb Ordering in English-Korean Machine Translation Using Clustering", Proceedings of nlprs2001, Tokyo, Japan, 2001.