결정트리를 이용한 중국어 복합명사 인식

최정[0] 김미영 김동일* 이종혁
포항공대 정보통신 대학원[0], 포항공대 전자 컴퓨터 공학부
첨단정보기술 연구센터
(cuizheng[0], colorful, dongil, jhlee)@postech.ac.kr

Recognition of Chinese Compound Noun Using Decision Tree

Cui Zheng[0] , Mi-Young Kim, Dong-il Kim*, Jong-Hyeok Lee
Graduate School of Information Technology0,
Dept. of Electrical and Computer Engineering,
Pohang University of Science and Technology,
And Advanced Information Technology Research Center (AITrc)

요 약

중-한 기계번역 시스템에 있어서 중국어 복합명사를 정확하게 처리하는 것이 매우 중요하다. 본 논문에서는 결정트리를 이용하여 품사의미 애매성을 포함된 복합명사를 인식하는 방법을 제시하였다. 제안한 방법으로 구축된 규칙이 기존의 방법보다 f-measure가 10.9% 더 좋은 성능을 보여 주었다.

1. 서 론

중-한 기계번역 시스템에서 중국어 복합명사를 정확하게 처리하는 것이 매우 중요하다. 복합명사가 정확히 처리되면 구 묶음 및 분석단계의 애매성을 줄여 줄 수 있다.

중국어는 무한한 조어능력을 가지고 있으므로 모든 가능한 복합명사를 사전에 등록할 수 없다. 또한, 똑같은 단어 쌍이 문맥에 따라 복합명사로 쓰일 수도 있고 그렇지 않을 수도 있으므로 복합명사를 인식하는 모듈이 필요하다.

아래 예의 2번 문장에서 기계번역(机器翻口)이라는 단어 쌍이, 두 개의 예문 중 1번 예문에서만 복합명사로 쓰였음을 알 수 있다. 따라서 기계번역(机器翻口)이라는 단어를 문맥을 고려하지 않은 채 복합명사로 인식하게 되면, 2번 문장이 아래의 3번 문장과 같은 잘못 번역된 결과를 가져오게 된다.

예 : 1. [**机器/vn 翻译/n**] 很/d 难/a 。/w
[기계 번역]은 아주 어렵다.
2. 他/r 用/p **机器/n 翻译/v** 文献/n. 。/w
그는 기계로 문헌을 번역한다.
3. 他/r 用/p [**机器翻口**]n 技口 /n 。/w
그는 기계번역 기술문헌을 쓴다.

따라서 본 논문은 문맥의 정보들을 사용하여, 결정트리를 기반으로 한 중국어 복합명사 인식방법을 제안한다.

여기에서 vn은 명사로 쓰일 수 있는 동사를 표시하고, n은 명사, r은 대명사, p는 전치사, d는 부사, w는 문장기호, a는 형용사를 표시한다.

2. 중국어 복합 명사 인식

기존 연구에서, Lixian Fan[1]은 중국어 복합명사 인식 규칙을 구축하고 애매성을 가지는 규칙들에 대해서는 제약조건을 추가 작성하였다.

다음 표는 복합명사 인식 규칙중의 일부이다.

규 칙	예
n+n+n	科学文化水平
n+vn	机器翻译
a+n	自然经济

<표 1> 복합명사 인식 규칙의 예

표 1에서 n+vn은 명사와 그 뒤의 동명사가 결합하여 복합명사를 구성한다는 규칙이다. 이 규칙에서 vn은 동사와 명사의 품사를 둘 다 가질 수 있는 단어를 의미한다. 이러한 동명사에 대한 품사 태깅의 정확률이 높지 못하므로(63%), 동명사가 나타나는

* 중국 길림성 연길시 연변과학기술대학 부교수

규칙에 대해서는 주변의 문맥을 고려한 더 많은 정보를 참조하여 복합명사인지 아닌지를 구별해야 한다. 동명사가 포함됨으로써 애매성을 가지는 복합명사 인식 규칙은 n+vn, vn+vn, vn+n 등이 있다.

Lixian Fan[1]은, 복합명사 인식 규칙의 애매성을 해소하기 위해 큰 규모의 말뭉치를 분석하여 좌우 문맥에 세 명사와 관계를 가지는 동사, 보조 동사, 동명사, 동사로 쓰일 수 있는 형용사 동 정보의 유무에 따라 규칙을 재구성하였다.

이 방법은 높은 정확성을 나타내는 반면 중국어 전문가가 말뭉치에 의거한 규칙에 대한 수작업을 요구하는 단점이 있기에 모든 경우를 고려하는 대규모 작업이 어렵게 되고, 재현율이 낮아지게 된다. 따라서 재현율을 더 높이기 위해서 다양한 규칙들을 구축할 필요가 생기게 된다.

2.1

앞에서 소개된 방법의 단점을 극복하기 위해 중국어 어법정보 사전[4][6]과 품사 태깅된 북경대학 인민일보 코퍼스[3]를 이용하여 자동적으로 규칙 구축하여 성능을 높일 시도를 하였다.

두 단어 쌍이 복합명사인지 아닌지 판단하기 위해 Lixian Fan이 제안한 규칙에 대해, 두 단어에 대한 제약조건을 좀 더 추가하고, 두 단어의 주변 문맥 정보를 좀더 사용할 필요성을 있다. 여기에서 두 단어에 대한 제약조건은 중국어 어법정보사전에서 추출하고, 두 단어의 주변 문맥 정보는 품사태깅된 코퍼스로부터 얻는다.

중국어 어법정보사전은 중국 북경대학에서 1986년부터 1996년까지 10년 동안 거쳐서 개발한 전자 사전이다. 이 사전에는 중국어 분석에 필요한 어법정보가 구축되어 있다. 어법정보사전을 통하여 애매성을 가진 복합명사패턴의 동명사의 속성에서 복합명사 인식에 도움을 줄 수 있는 속성들을 추출한다.

품사 태깅된 인민일보 말뭉치에는 복합명사 또한 표시되어 있다. 이 말뭉치로부터 복합명사의 앞뒤 품사 정보를 얻을 수 있다. 아래는 문맥정보를 이용한 예이다.

예: 机器/n 翻译/vn 了/u 外国/n 的/u 文献/n 。/w

기계는 외국의 문서를 번역하였다.

위의 예에서 n+vn 의 뒤에 조사가 오면 이 패턴은 복합명사가 아니라는 규칙을 얻을 수 있다. 이는 문맥정보가 복합명사 추출에 도움이 된다는 증거이다. 문맥정보를 이용한 이런 규칙들은 품사태깅된 코퍼스로부터 자동으로 학습할 수 있다.

표2는 규칙을 재구축하는데 사용할 속성들의 설명이다.

Feature		
后名 backn	뒤에 명사와 결합가능 여부 (가능 = 1, 불가 = 0)	翻译(backn=1) 文献 번역 문헌
助动 auxv	규칙 앞 단어 조동사 여부 (조동사 = 1, 기타 = 0)	跑 去(助动) 달려 가다
趋向 dirv	규칙 앞 단어 방향동사 여부 (방향동사 = 1, 기타 = 0)	走 上(趋向) 위로 가다
兼语 pivv	규칙 앞 단어 피봇동사 여부 (피봇 동사 = 1, 기타 = 0)	帮(兼语) 他 打扫 그를 도와 청소하다
有宾 objv	규칙 앞 단어 타동사 여부 (타동사 = 1, 기타 = 0)	看(有宾) 电影 영화를 보다
앞 품사 frontpos	규칙에 만족하는 패턴 앞 단어의 품사	
뒤 품사 backpos	규칙에 만족하는 패턴 뒤 단어의 품사	

<표 2> 결정트리에서 사용할 속성들

2.2 결정트리를 이용한 학습

우리의 방법은 Lixian Fan[1]의 복합명사 인식 규칙들을 기본으로 사용하고 동명사를 포함한 규칙들에 대해서는 표 2에서 제시한 속성들을 사용하여 인식 정확성을 높인다.

복합명사가 표시 되어 있는 품사 태깅된 말뭉치로부터, 복합명사가 가능한 품사패턴 (n+vn, vn+vn, vn+n,)을 추출하고, 결정트리를 이용하여 이 패턴들에서 복합명사 인식 규칙을 얻는다.

결정트리(Decision tree) 기반의 방법으로 Quinlan[2]의 C4.5 프로그램이 널리 쓰인다. C4.5는 엔트로피에 근거하여 루트 노드에서 단말 노드순으로 트리를 구성해 나간다. 각 노드들에서는 이진 비교가 수행되며 단말 노드에 이르면 분류 결정이 완료된다. 이런 결정 트리는 그 결과의 가독성 및 해석력이 우수하기 때문에 널리 사용되고 있다.

엔트로피[5]는 기계학습 분야에서 특성의 영향력 측정 수단으로 널리 쓰여 왔다. 이것은 데이터 내에 특성 값의 유무를 알게 됨으로써 분류를 위해 얻어지는 정보량의 비트수를 측정한다. 특성 값 t에 대한 정보 이득은 다음과 같이 정의된다.

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i \mid t)\log P_r(c_i \mid t) + P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i \mid \bar{t})\log P_r(c_i \mid \bar{t})$$

$\{C_i\}_{i=1}^{m}$ 은 클래스의 집합이다. 주어진 학습 데이터 집합에 대하여 각 특성에 대해 정보 이득을 계산 후, 임계치 미만의 특성들을 제거하는 것이다.

3. 실험 및 평가

말뭉치에서 애매한 패턴들에 대해 결정트리 생성 프로그램 C4.5로 규칙을 추출한다. 아래는 그 과정에 대한 설명이다.

품사 태깅된 코퍼스에서 n+vn, vn+vn, vn+n 패턴들을 찾고 그 동사의 중국어 어법정보사전으로부터 2.1에서 설명한 정보들을 추출한다. 그리고 그 패턴의 좌, 우 단어의 품사 정보도 추출하고 이 패턴이 복합명사이면 class 2로 지정하고 아니면 class 1로 지정하여 2000개의 훈련데이터와 1000개의 평가 데이터를 추출한다.

다음 C4.5 프로그램으로 훈련데이터에서 규칙을 추출하고 실험 평가 데이터로 추출된 규칙들의 성능을 평가한다.

아래는 C4.5로 훈련시킨 뒤 추출된 규칙의 예이다.

예 3: Rule 7: frontpos = d
→ class 1
Rule 8: backpos = n
backn = 1
→ class 2

Rule 7은 패턴 앞 단어의 품사가 d(부사)이면 class 1(즉 복합명사로 될 수 없다)이 되는 규칙이고 Rule 8은 패턴 뒤의 품사가 n(명사)이고 동명사가 명사와 결합이 가능하면 복합명사로 된다는 의미이다.

같은 훈련 데이터와 테스트 데이터에 대한 결정트리로 진행한 중국어 복합명사 인식 실험에 대해 설명해 본다.

실험 1: Lixian Fan[1]의 방법으로 한 성능을 평가

실험 2: 실험 1의 정보에 단어 쌍의 어법정보 추가

실험 3: 실험 2의 정보에 규칙의 좌우 문맥에서 인접한 각각 한 단어의 품사정보 추가

실험 4: 실험 2의 정보에 규칙의 우변 문맥에서 두 단어의 품사정보 추가

실험 5: 실험 2의 정보에 규칙의 좌변 문맥에서 각각 두 단어의 품사정보를 추가

실험 6: 실험 2의보에 규칙의 좌우 문맥에서 각각 두 단어의 품사정보 추가

표 3는 각 실험에 대한 성능평가 결과이다. 실험 1 즉 Lixian Fan[1]의 방법이 좋은 성능을 보여주지 못한 원인을 분석해 본 결과, 수작업으로 작성된 룰의 적용률이 아주 낮다는 것을 발견하였다. 현재 사전에 들어 있는 어법 정보의 빈약으로 실험 2의 방법이 좋은 성능을 보여주지 못 하였다.

실험 3~실험6의 결과들이 문맥 정보를 이용한 결정트리로 아주 좋은 성능을 보여줬다.

	Compound Recognized	Correct	Precision	Recall	F-Measure
실험 1	528	374	70.8%	68.37%	69.5%
실험 2	598	396	66.2%	72.3%	69.11%
실험 3	490	377	76.93%	68.9%	72.69%
실험 4	550	411	74.7%	75.1%	74.89%
실험 5	613	466	76.01%	85.1%	80.2%
실험 6	613	467	76.1%	85.3%	80.4%

<표 3> 중국어 복합명사 인식 실험 결과 비교

실험3과 실험 4에서 패턴 앞뒤 단어의 품사정보가 복합명사 인식에 도움이 확실하다는 사실을 확인할 수 있다. 실험 5와 실험 6에서 더 많은 문맥이 복합명사 인식에 더 많은 정보를 주는 것을 알 수 있다. 실험 5에서 패턴 앞의 문맥정보가 복합명사인지를 결정하는데 충분한 정보를 준다는 결론을 얻을 수 있다.

4. 결론 및 향후 작업

본 논문에서 자동으로 중국어 복합명사 인식 규칙을 구축하는 방법을 제시 하였다. 실험에서 보여주는 바와 같이 좌, 우 두 단어의 품사, 그리고 경용품사의 속성으로 구축한 규칙으로 인식한 결과의 f-measure값이 기존의 방법보다 10.9% 높은 성능을 보여 줬다. 향후 진행해야 할 연구는 오류에 대한 상세한 분석과 다른 기계학습(machine learning)기법에서 정확성을 더 높일 수 있는 방법을 제시하는 것이다.

5. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

6. 참고문헌

[1] Linxin Fan, Fuji Ren, Yoshikazu miyanaga and Koji Tochinai, "Automatic Composition of Chinese Compound Words for Chinese-Japanese Machine Translation," in Transactions of Information Processing Society of Japan, Vol. 33, No.9, 1992.

[2] Quinlan, J. R., "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers Inc., 1971.

[3] Yu Shiwen, Zhu Xuefeng, Duan Huiming, "The Guideline for Segmentation and Part-Of-Speech Tagging on Very Large Scale Corpus of Contemporary Chinese," in the 2000' International Conference on Multilingual Information Processing. (00' ICMIP).

[4] Yu Shiwen, Zhu Xuefeng, Wang Hui, "New Progress of the Grammatical Knowledge-base of Contemporary Chinese," in Journal of Chinese Information Processing, Vol.15, No.9, 2001.

[5] 김영택, "자연언어처리," 생능출판사, 2001.

[6] 北京大学计算语言学研究所, "《现代汉语语法信息词典》规格说明书," 2000.