백과사전 질의응답을 위한 생략된 표제어 복원에 관한 연구

임수종° 이창기 장명길
한국전자통신연구원 음성/언어정보연구부 지식마이닝연구팀
{isj, leeck, mgjang}@etri.re.kr

Restoring an Elided title for Encyclopedia QA System

Soojong Lim°, Changi Lee, Myoung-Gil Jang
Speech/Language Information Research Department, ETRI

요 약

백과사전에서 정답을 찾기 위해 문장의 구조를 분석하는데 한국어 백과사전은 표제어에 대한 정보를 문장에서 생략한다. 그러나 표제어는 문장에서 주어나 목적어 역할을 하기 때문에 생략된 정보를 복원하지 못 하면 질의에 대한 정답을 제시할 수 없다.

생략된 표제어에 대한 정보를 복원하기 위해서 본 연구에서는 표제어의 의미범주 정보, 격틀, Maximum Entropy 모델을 이용하여 표제어 주어, 표제어 목적어 복원, 미복원 3가지로 인식한다. 표제어 의미범주는 의미 범주에 대해 일정 수준의 복원 성향을 보일 경우 Maximum Entropy 정보를 참조하였고 격틀을 이용하여 복원 여부를 결정한다. 만약 표제어의 의미범주 정보, 격틀을 이용하여도 복원 여부를 결정하지 못할 경우에는 Maximum Entropy 모델에 기반한 통계 기법을 적용하여 복원 여부를 결정한다. 그리고 각각 방법의 단점을 보완하기 위해서 규칙에 해당하는 표제어 의미범주 정보와 격틀 정보에는 통계 모델인 ME 모델을 보완하여 사용한다.

1. 서론

자연언어 처리에서 발생하는 생략(ellipsis)은 주로 반복을 피하기 위해서 발생하기 때문에 같은 문장이나 혹은 그 전의 문장에서 유추하여 생략 현상을 복원할 수 있다[4]. 이러한 일반적인 문장에서의 생략 현상에 대해서는 많은 연구가 진행되었지만 다음과 같이 특정한 언어적 현상과 응용 시스템에서 발생하는 현상에 대한 연구가 필요하다.

영어의 경우 생략 현상은 다음과 같이 동사구가 생략된 형태가 된다.

Helen saw the movie and Mary did too.

한국어의 경우에는 동사구가 생략되는 형태는 극히 드물고 주로 주어나 목적어 역할을 하는 명사구가 생략이 된다. 이러한 현상은 질문에 대해 정확한 답만을 제시하는 질의 응답 시스템에서도 발생되는데, 질의응답 시스템에서 답을 주기 위해 여러 가지 기법이 사용되는데 그 중의 하나로 다음과 같이 질문과 정답 문장의 술어-논항 관계를 이용하는 방법이 있다.

2000년에 노벨평화상을 받은 사람은?
받다(subj:사람, obj:노벨평화상, adv:2000년)

Title: 김대중
...공로로 2000년 노벨평화상을 받았다.
받다(subj:김대중, obj:노벨평화상, adv:2000년, 공로) (1)
받다(subj:NULL, obj:노벨평화상, adv:2000년, 공로) (2)

생략 현상이 빈번하게 일어나는 한국어 백과사전 문서는 실제 문장 안에서 표제어에 해당하는 부분은 모두 생략을 한 상태로 문장이 구성되기 때문에 문장 그대로 술어 논항 관계를 구성한 경우 (2)와 같은 형태가 되어 원하는 정답을 찾을 수가 없다. 본 연구의 목적은 이러한 생략현상을 극복하여 (1)과 같은 형태로 표제어를 복원하는 것이다.

일반적으로 생략된 요소를 복원하는 것은 선행하는 문장의 요소 중에서 반복을 피하기 위해 생략한 요소를 복원하는 것이지만 백과사전 표제어는 어느 문장에서도 사용된 적이 없기 때문에 기존의 문제와는 조금 다른 양상을 띠게 된다. 표제어는 문장 구조에서 주어와 목적어일 경우 생략이 되고 이러한 표제어를 복원하는 것은 동일 문장에서 반복을 피하기 위해 생략되는 형태와 일반 주어의 형태로 복원할 필요가 없는 경우와 구분을 하여 복원을 하여야 한다. 예를 들자면, 문장에서 주어가 생략됐다고 해서 무조건 복원을 하여서는 안 된다.

본 연구는 백과사전 질의 응답 시스템에 술어-논항 관계를 이용하여 정답을 찾기 위해 사용되는 문장 구조 분석기의 일부분으로 사용된다. 표제어의 복원 여부를 결정하기 위해서 표제어에 대한 의미 코드를 이용하여 주어로 사용되며 잘 생략되는 사람 등과 같은 의미코드를 사용하고 구조 분석된 문장을 격틀처럼 변형하여 격틀 사전에서 필수격으로 사용되며 생략된 형태가 표제어의 의미코드와 같은지 여부를 조사하여 매칭이 되는 격틀이 있을 경우 복원하도록 하였다. 그러나 두가지 방법의 단점을 보완하고 두 가지 경우로 처리하지 못 하는 경우에 대해서는 용언의 품사, 연결어미, 용언간의 의존관계, 표제어의 의미코드 등을 자질으로 하여 Maximum Entropy(ME) 모델을 학습한 후에 적용하여 표제어 복원 여부를 판단하는 방법을 사용하였다.

2. 관련 연구

생략 현상은 언어처리에서 계속되어진 문제이고 관련 연구가

문장의 구조와 의미 표현이라는 측면에서 많이 진행되었지만 실제적으로 언어공학적 관점의 관련 연구는 그리 많지 않다.

관련 연구로는 첫번째 Penn Treebank를 이용하여 동사구(VPE) 생략을 찾아내는(detection)는 것에 국한한 연구[2]가 있다. 정확률 44%, 재현률 53% F-measure 48% 정도의 성능을 보였으나 단순한 탐색기법을 사용하여 생략 현상을 찾아내는 것에만 한정을 지어서 실제적으로 어떤 동사를 복원해줘야 하는지에 대한 언급이 없다. 둘째로는 기계학습 기법을 사용한 연구[5]이다. TBL(Transformation-based learning), ME(Maximum entropy modeling), Decision Tree Learning, MBL(Memory Based Learning) 등 기계학습 기법을 사용하였다. 기계학습을 사용한 연구의 경우 정확률 85.14%, 재현률 69.63%, F-measure 76.61%의 결과를 보였으나 앞의 연구와 똑같이 생략 여부만을 연구대상으로 했기 때문에 본 연구와 직접적으로 비교를 할 수가 없다.

본 연구는 각각 규칙과 통계를 사용한 방법을 적용하여 보고 각각 접근 방법의 장단점을 파악하여 장점만을 결합하여 규칙과 통계를 이용한 방법을 사용한다.

3. 표제어 복원

제안하는 방법은 규칙과 통계를 복합적(hybrid)으로 사용하는 방법으로 백과사전의 표제어 중에서 많은 부분을 차지하는 의미 코드 중에서 주로 주어나 목적어 중에서 빈번히 생략되는 의미코드를 선정하여 무조건 복원하도록 하고 이외의 경우에는 격틀을 이용하여 정밀하게 복원 여부를 결정한다. 2가지 경우를 보완하기 위해서 통계기법인 ME 모델을 학습하여 적용하는데 표제어의 의미코드를 사용할 경우에는 무조건 복원되는 위험성을 보완하기 위하여 ME 모델에서 복원되지 않을 경우에 대해 threshold를 설정한다. 그리고 격틀에 대해서는 격틀에 존재하지 않아서 판단하기 힘든 경우에 대해서 판단을 하게 된다.

3.1 표제어 의미코드를 이용한 방법

격틀은 400여개의 의미 코드를 사용하여 구성되어 있다. 이러한 의미 코드를 이용하여 백과사전의 10만여 문서의 표제어에 대해 백과사전의 범주 정보와 정의문을 사용하여 의미 코드를 반자동으로 할당하였다. 의미코드가 할당된 문서 중에서 같은 의미코드가 200개 이상인 75개의 의미코드를 대상으로 하여 각 문장에서 표제어의 생략 정도를 수작업으로 조사하였다. 총 36개의 의미코드가 50% 이상의 용언에 대해서 표제어가 주어로 사용될 경우 생략이 되서 이러한 의미코드는 복원대상 의미코드로 설정하였고 목적어에 대해서는 최고 40% 미만의 확률을 생략이 되어서 무조건 복원하는 의미코드로 채택하기에는 적합하지 않아서 제외하였다. 대상 의미코드는 다음의 [표1] 같다.

표 1 복원대상 의미코드

사람, 집단, 학문, 창작물, 곳, 동물, 예술, 식물, 건축물, 업무, 지위, 운동경기, 의료, 시설, 출판물, 장치, 교통기관, 현상, 하드웨어, 방송, 인종, 존재, 갈래, 물체, 무기, 폭발물, 언어, 시설물, 동작, 기호, 지형, 길, 경제, 광고, 사상, 무덤

3.2 격틀

백과사전 문장을 구조 분석하기 위한 문장 구조 분석기[5]은 격틀에 기반하여 문장을 구조 분석한다. 격틀은 다음과 같은 형태로 구성된다.

A=의미코드! 격조사 용언!다 > 예문

A=사람!가 B=인공적장소!로 가!다 > [그[A]가 바다[B]로 가다]

3만여개의 용언, 약 15만3천여개의 격틀을 사용하였다.

의미코드는 ETRI 명사개념망[1]에서 상위 노드 444개를 정하여 사용하였다. 의미코드에 대한 정의는 표준국어대사전을 이용하였다.

격틀을 이용한 방법은 문장구조 분석기의 결과를 이용하여 격틀에 매칭을 해 본 후에 표제어로 생략된 주어나 목적어를 제외한 격틀이 존재할 경우에는 복원 여부를 결정한다. [표2]는 격틀을 이용하여 표제어를 주어로 복원하는 여부를 결정하는 예이다.

표 2 격틀을 이용한 표제어 복원 예

입력	Title: 알롱만(Along Bay) sense: Location Sentence: 하이퐁 동쪽에 위치하며
문장구조분석(parsing)	위치하다(subj:NULL, obj:NULL, adv: 동쪽!에) 격틀: 방향!에 위치하다
격틀 매칭	24265-2 A=곳!가 B=곳!에서 C=*방향!에* 24265-4 A=곳!가 B=*방향!에* 24265-8 A=기상!가 B=*방향!에* 24265-12 A=*방향!에* 24265-17 A=부위!가 B=*방향!에*
복원여부	주어로 복원

3.3 ME 모델을 이용한 통계 방법

규칙과 격틀 기반 방법으로 결정하기 힘든 문제를 통계 정보를 사용하고자 해결하고자 한다. 최대 엔트로피(Maximum Entropy, ME) 모델은 주어진 제약 조건을 만족하는 여러 확률 분포 중에서 가장 균일한 분포 상태를 가지는 모델이다. 바꾸어 말하면, ME 모델은 주어진 제약 조건 하에서 최대 엔트로피를 가지는 확률 분포를 가지고 있다. 이를 수식으로 나타내면 다음과 같다.

$$P = \{ \text{models consistent with constraints} \}$$
$$H(p) = \text{Entropy of } p,\ p \in P$$
$$P_{ME} = \text{argmax}_{p \in P} H(p)$$

여기서 P_{ME} 가 최대 엔트로피 확률 분포를 가지는 모델이다.

ME 모델의 매개변수 추정에 사용되는 알고리즘에는 Generalized Iterative Scaling(GIS), Improved Iterative Scaling(IIS), 그리고 Limited Memory BFGS(L-BFGS) 등 잘 알려진 것이 몇가지 있다. 본 연구에서는 GIS 알고리즘을 사용하였다.

ME 모델의 가장 두드러진 특징은 모델의 특성을 완전히 드러내는 후보 자질들을 선택해 주기만 하면 되는데 본 논문에서는 후보 자질로 어휘자질(lexical feature), 품사자질(POS feature), 의미자질, 구조분석 자질, 복합자질 등을 사용하였다. 각 자질은 다음과 같다.

[1] ETRI에서 구축중인 개념망으로 명사, 동사 개념망으로 구성되고 명사는 약 6만여개의 노드가 있다.

Verb_lex : 표제어 복원 대상 용언의 어휘
Verb_pos : 표제어 복원 대상 용언의 품사
Verb_e_lex : 표제어 복원 대상 용언에 부착된 어미의 어휘
Verb_e_pos : 표제어 복원 대상 용언에 부착된 어미의 품사
Ti_res_code: 표제어 의미코드를 이용한 대상에 포함되는지 여부
Verb_cf_subj, obj: 표제어의 의미코드가 용언의 격틀에서 주어 혹은 목적어로 사용되는지 여부
Ti_sense: 표제어 의미코드
Tree_posi: 문장구조분석 트리에서 상위, 중간, 하위로 구분
Rel_type: 문장구조 분석 트리에서 다른 용언이 존재할 경우 어미를 이용한 용언간 관계.
Sen_subj, obj: 구문 분석 결과 용언에 대해 주어 혹은 목적어가 존재하는지 여부
Pair: 표제어 의미코드와 용언 쌍

3.4 생략된 표제어 복원 알고리즘

복원 여부를 최종적으로 결정하기 위해서는 기본적으로 첫째, 표제어의미코드, 둘째, 격틀, 셋째 ME 모델을 순차적으로 이용하여 복원 대상으로 되면 다음 단계로 넘어가지 않고 알고리즘을 멈춘다. 그러나 순차적으로 적용을 할 경우에는 뒤쪽의 방법을 사용할 경우 더 정확한 결과를 얻을 수 있는 경우가 앞쪽에서 판단을 내려서 원하지 않는 결과가 될 수 있다. 이것은 3가지 방법은 서로 다른 특성이 있기 때문인데 예를 들자면 표제어 의미코드를 사용하면 무조건 복원을 하기 때문에 기본적으로 오류를 포함하고 있고 해당되지 않는 의미코드에 대해서는 복원을 할 방법이 없다. 정확한 격틀을 사용할 경우 복원 정확률은 다른 방법에 비해 상대적으로 높지만 재현률이 떨어지는 문제가 있다. 이러한 특성에서 단점을 보완하기 위해서 다음과 같은 단계를 거친다.

첫번째 표제어 의미코드를 사용하는 방법의 단점인 복원에 해당하지 않는 표제어에 대한 복원과 복원 대상에 대한 상대적으로 낮은 정확율을 보완하기 위해서 ME 모델을 결합하였다. 표제어 의미코드에 해당하여 복원 대상이 되더라도 ME 모델을 사용하여 부정적인 임계값(negative threshold)을 넘는 경우라면 복원하지 않는다.

두번째, 격틀을 이용하는 방법을 보완하기 위해서 ME 모델을 결합한다. 격틀을 이용할 경우 1단계로 완전 매칭(exact match)을 사용하지만 이 경우는 정확률은 높지만 재현율이 낮다. 2단계로 부분 매칭(partial match)을 사용하면 재현율은 상대적으로 높아지지만 정확률은 낮아질 수 있기 때문에 이 경우에 앞서와 같은 방법으로 ME 모델을 적용하여 정확률을 보완한다.

4. 실험 및 분석

ME 모델을 학습하기 위한 실험데이터를 수동으로 구축하였다. 실험 데이터는 백과사전에서 표제어의 의미코드를 고려하여 추출된 문장에서 표제어 주어 복원(916용언), 표제어 목적어 복원(223용언), 미복원(1756용언)의 3가지 분류로 총 2895 용언에 대해 구축하였다.

평가는 격틀만을 사용한 실험(CF), 표제어 의미코드만을 사용한 실험(ASC), ME 모델만을 적용한 실험(ME), 표제어 의미코드와 격틀을 사용한 실험(ASC_CF), 표제어 의미코드와 ME 모델을 사용한 실험(ASC_ME), 표제어 의미코드, 격틀, ME 모델을 사용한 실험(ASC_CF_ME)로 나누어 실행하였다. 평가 데이터는 총 277문장으로 평가에서 baseline은 표제어를 무조건 주어로 복원하는 경우로 하였다. 실험결과는 표[3]과 같다.

표 3 실험결과

	Recall	Precision	F-measure
Baseline	100.00	31.64	48.07
ASC	58.14	66.37	61.98
CF	56.91	56.45	56.68
ME	62.50	35.40	43.52
ASC_ME	79.03	59.39	67.82
ASC_CF	68.55	50.00	57.82
ASC_CF_ME	78.23	60.25	68.07

표제어 의미코드는 복원 확률이 50% 이상인 경우만을 대상으로 하기 때문에 높은 정확률을 보였지만 정해진 의미 코드에 해당하는 표제어만 처리할 수 있기 때문에 상대적으로 낮은 재현률을 보인다. 격틀 방법은 완전 매칭과 부분 매칭을 함께 사용하였기 때문에 재현율과 정확률이 비슷한 결과를 보였다. 세가지 방법을 결합할 경우 제안하는 방법과 같이 가장 좋은 결과를 나타낸다. 이러한 실험 결과는 제안하는 방법이 의도한대로 각각의 방법의 단점이 다른 방법에서 효과적으로 수정되어져 올바른 결과를 냈다는 것을 알 수 있다.

5. 결론 및 향후 연구방향

본 연구에서 백과사전 질의응답 시스템에서 정답을 찾기 위해서 문장 구조 분석을 할 때 필요한 정보 중의 하나인 표제어 복원 여부을 판단하기 위해서 표제어의 의미코드, 격틀, ME 모델을 이용하고 이들의 단점을 보완하기 위해서 혼합하여 사용하는 알고리즘을 제시하였다. 각각의 방법만을 사용하였을 경우 각 접근 방법의 단점에 있어서 성능은 60을 넘지 못하는 수준이지만 제안하는 방법과 같이 각 방법의 단점을 상호 보완해줄 경우 F-measure 68.07을 보였고 비슷한 문제인 한국어의 명사 생략에 적용했을 때도 비슷한 결과를 보였다.

향후 연구로는 첫째 ME 모델 이외에도 다른 기계학습 방법을 적용하여 비교하여 보고 한국어 생략 현상에 가장 적합한 기계학습 모델을 찾는 것이고 둘째는 앞에서도 언급했듯이 백과사전의 표제어 생략이라는 특정한 문제 해결을 목표로 접근한 방법을 좀더 일반적인 생략현상에도 적용할 수 있도록 알고리즘을 개발하는 것이다. 마지막으로 상대적으로 낮은 성능을 보인 목적어를 복원하는 알고리즘을 찾는 것이다.

6. 참고 문헌

[1] Adwait Ratnaparkhi. *Maximum Entropy Models for Natural LANGUAGE Ambiguity Resolution*, Unpublished PhDthesis, University of Pennsylvania. 1998.
[2] Daniel Hardt. *An empirical approach to vp ellipsis*, Computational Linguistics, 23(4). 1997.
[3] H. J. Kim, H. J. Oh, C. H. Lee., et al. *The 3-step Answer Processing Method for Encyclo-pedia Question-Answering System: AnyQuestion 1.0.* The Proceedings of Asia Information Retrieval Symposium (AIRS) 309-312 2004.
[4] James Allen. *Natural Language Understanding*, Benjamin/Cummings Publishing Company, 449~455 1995.
[5] Leif Arda Nielsen. *Using Machine Learning Techniques for VPE detection*, RANLP 03, Bulgaria. 2003.
[6] Lim soojong. *Dependency Relation Analysis Using Caseframe for Encyclopedia Question-Answering Systems*, IECON, Korea. 2004.