일반화된 패턴을 이용한 관계 추출 시스템

김혜민° 최익규 김민구
아주대학교 정보통신전문대학원°, 아주대학교 정보 및 컴퓨터 공학부
{khm°, ikchoi, minkoo }@ajou.ac.kr

Relation Extraction System using Generalized Patterns

Hyemin Kim°, Ikkyu Choi, Minkoo Kim
Graduate School of Information and Communication, Ajou University
College of Information & Computer Engineering, Ajou University

요 약

텍스트 형태의 문서에는 많은 종류의 유용한 관계가 존재한다. 이러한 관계들을 문서에서 자동으로 찾아내는 것은 정보검색 분야에서 매우 중요한 작업 중 하나이다. 그러나 각각의 관계마다 다양한 형태의 패턴이 존재하기 때문에 많은 양의 문서에서 이러한 관계들을 찾아 내는 것은 쉬운 일이 아니다. 이러한 어려움을 해결하기 위해 본 논문에서는 일반화된 패턴을 이용하여 자동으로 관계를 찾는 방법을 제안한다. 본 논문에서 제안하는 방법은 초기에 사용자로부터 얻은 정보를 이용하여 관계를 자동으로 찾는다. 약 1,000,000개의 문장을 이용해 실험한 결과 자동으로 일반화된 패턴을 이용하는 방법을 이용할 경우 그렇지 않은 경우보다 성능이 향상됨을 확인할 수 있었다.

1. 서 론

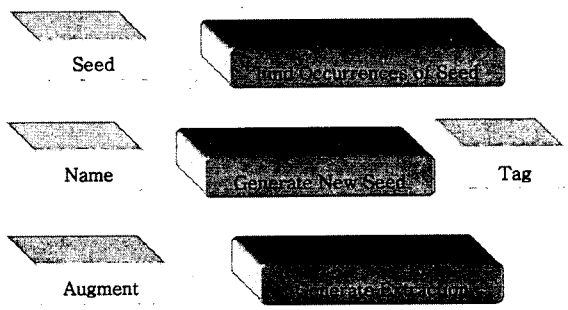
실 세계는 개념이라는 단위로 구성되어 있으며 개념간에는 수 많은 관계가 존재한다. 이러한 관계들은 text형태의 문서에서 다양한 형태로 나타나게 된다. 인터넷 기술의 급격한 발전으로 문서의 수도 많아지게 되었고 많은 양의 문서에서 자동으로 관계를 찾는 방법에 대한 연구[1,3,5]도 활발하게 진행되고 있다. 자동으로 관계를 찾는 시스템의 핵심은 실제 문서에서 관계를 추출할 때 쓰이는 패턴을 생성하는 것이다. 패턴은 관계가 실제 문서에서 실현되는 모습을 말한다. 예를 들어, " X 는 Y이다." 라는 문장에서 " ~는 ~이다." 라는 패턴은 상, 하위 관계가 문장에서 실제 표현된 형식이라고 할 수 있다. 관계 추출 시스템은 이러한 패턴을 이용하여 문서에서 자동으로 관계를 추출할 수 있다. 만들어진 패턴이 정확할수록 추출된 관계 정보도 정확해진다.

자동 관계 추출 시스템 중 Snowball[5]이라는 시스템은 사용자가 주는 초기 정보를 이용하여 패턴을 만들고 관계를 자동으로 추출한다. 사용자가 주는 정보를 근거해서 패턴을 추출하기 때문에 다른 방법보다 유연하다고 할 수 있다. 그러나 이 방식은 만들어진 패턴을 가공하지 않고 그대로 사용하기 때문에 다양한 패턴을 모두 수용할 수 없었다. 본 논문에서는 다양한 패턴을 수용하기 위해 SP+PR[6]F에서 이용한 방법을 적용하여 일반화된 패턴을 이용해 상, 하위 관계를 추출하는 방법을 제안하였다.

본 논문에서는 2장 관련 연구에서 Snowball System을 살펴보고, 3장에서는 본 논문에서 제안한 방법을 설명한다. 4장에서 테스트 환경 및 결과를, 5장에서 결론 및 향후 연구 과제를 기술한다.

2. 관련 연구

Snowball은 초기에 사용자가 입력하는 적은 양의 정보를 이용하여 장소와 기관간의 관계를 자동으로 찾는 시스템이다. Snowball에서는 장소와 기관을 표시하기 위해 named-entity tagger를 이용하였다.



[그림 1] Snowball 시스템의 흐름

본 논문은 산업자원부의 국가지정연구실 사업의 일환으로 지원 받아 수행되었음. (과제명:차세대 인터넷을 위한 지능형 온톨로지 자동생성 시스템 개발, 과제번호 M10302000087-03J0000-04400)

2.1 Snowball 시스템

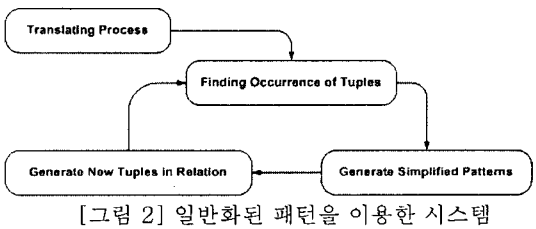
Snowball에서는 사용자로부터 초기에 유효한 장소와 기관으로 이루어진 쌍을 입력 받고 문서상에서 이 seed가 나타난 부분을 vector로 만든다. 예를 들어, 장소를 tag1, 기관을 tag2라고 한다면 문서에서 장소와 기관이 나타난 부분을 5-tuple, $t=<l_c,tag_1,m_c,tag_2,r_c>$와 같은 형태로 나타낸다. lc는 tag1의 왼쪽 컨텍스트, mc은 tag1과 tag2의 사이에 있는 컨텍스트, rc는 tag2의 오른쪽의 컨텍스트를 말한다. Vector의 각 요소의 weight는 해당 term의 빈도수에 따라 매겨지게 된다. 이러한 vector를 클러스터로 만들어 클러스터의 centroid를 패턴으로 사용하게 된다. 패턴이 만들어 지면, 문서에서 장소 *o*와 기관*l*이 함께 나타난 컨텍스트를 5-tuple로 만들어 위에서 만들어진 패턴과 비교하여 유사도 $Match(t,t_p)\geq\tau_{sim}$일 경우 *o*와 *l*을 장소와 위치 관계에 있는 새로운 쌍이라고 판단하게 된다. Snowball에서 장소-위치 관계에 있는 새로운 쌍을 찾는 데에 중요한 역할을 하는 것은 클러스터의 centroid인 패턴이다. 그런데 이 패턴은 문장에서 장소와 기관 사이에 나오는 모든 단어를 이용하여 만들어지기 때문에 의미 없는 단어가 포함되어 다양한 종류의 패턴을 인식하지 못하는 문제가 생긴다.

2.2 SP+PRF

SP+PRF System은 단어에 대한 정의를 찾아주는 QA System이다. 이 시스템은 기존의 Knowledge Base에 존재하지 않는 단어에 대한 정의를 관련된 신문 기사에서 찾아주는 시스템이다. 본 논문에서는 일반화된 패턴을 만들기 위해 SP+PRF 시스템에서 사용한 방법을 적용하였다.

3. 일반화된 패턴을 이용한 관계 추출 시스템

제안된 방법에서는 2장에서 언급한 문제를 해결 하기 위해 일반화된 패턴을 이용하였다. 본 논문에서는 개념들의 가장 기본적인 관계[2]중의 하나인 상, 하위 관계를 찾는 데에 초점을 맞추었다. Snowball이 기관과 장소를 찾는 데 named-entity tagger를 사용한 반면, 본 논문에서는 상, 하위 관계에 있는 단어를 찾기 위해서 C-value와 NC-value[8]를 이용하였다.



[그림 2] 일반화된 패턴을 이용한 시스템

3.1 Translating Process

본 논문에서 제안한 방법에서는 일반화된 패턴을 생성하기 위해서 문장의 핵심이 되는 부분만을 사용하였다. 따라서 형용사와 부사, 기타 수식어들은 제거하고 일부 단어들은 특정 분류로 구분되어 대표 단어로 바뀌게 된다. 품사를 구분하기 Brill Tagger[7]를 사용하였다. Translation Process에서 사용한 규칙은 아래의 표와 같다. 표1은 SP+PRF에서 사용한 방법을 단순화 시킨 것 이다. SP+PRF에서 사용한 규칙과 본 논문에서 사용한 방법 간의 가장 큰 차이는 제안된 방법에서는 명사와 명사절을 그대로 사용한다는 것이다. 상, 하위 관계의 주를 이루는 단어가 명사와 명사절이기 때문에 이들은 변형되거나 제거되지 않고 그대로 사용된다.

Terms/POS tag	Category	Example
is, are, am, was, were	BE	Is ->BE
Noun, Noun phrases	Not translated	Kodak-> Kodak
Adjectival and Adverbal modifiers	To be deleted	
Determiner	DT	the ->DT

[표 1] 단어의 분류 및 변형 규칙

3.2 일반화된 패턴의 생성

문서에서 상, 하위 관계에 있는 단어들의 쌍을 찾기 위해 초기에 사용자로부터 적은 양의 seed의 쌍을 입력 받는다.

상위어	하위어
Protocols	Transmission Control
Program	Gauss
Telecommunication company	Bell
Firm	Insignia
Software	Network Computing System

[표 2] 상, 하위 관계를 위한 초기 seed의 예

표 2의 단어들이 문서에 나타난 부분의 컨텍스트를 묶어 1장에서 설명한 것처럼 5-tuple을 만들고 single-pass 클러스터 알고리즘을 이용해 클러스터를 만들게 된다. 이 때 사용되는 Match function은 아래와 같다.

$t_p=<l_p,tag_1,m_p,tag_2,r_p>$ 이고 $t_s=<l_s,tag_1,m_s,tag_2,r_s>$ 일 경우

$$Match(tp,ts)=\begin{cases} l_p\cdot l_s+m_p\cdot m_s+r_p\cdot r(\text{if the tags match}) \\ 0\ (\text{otherwise}) \end{cases}$$

이 과정에서 생성된 패턴 즉, 클러스터의 centroid의 예는 표 3과 같다.

Left vector	Middle vector	Right vector
<hardware, 0.354> <and, 0.354> <support, 0.354>	<DT,0.604> <systems, 0.250> <including, 0.604> <server,0.250>	<software, 0.354> <protocol, 0.354> <tcp/ip, 0.354>

[표 3] 일반화된 패턴의 예

3.3 상, 하위 관계에 있는 새로운 seed의 생성

새로운 상, 하위 관계에 있는 단어의 쌍을 찾기 위해 NC-value와 C-value를 기준으로 선정된 단어들이 t1, t2가 나온 컨텍스트를 모두 묶어 5-tuple t을 만든다. Weight를 주는 방법은 1장에서 설명한 Snowball의 방식과 동일하다. 이때 만들어진 t와 앞 장에서 만들어진 클러스터의 centroid p간의 유사도 $Match(t,t_p) \geq \tau_{sim}$ 일 경우 t1, t2를 상, 하위 관계에 있는 새로운 쌍이라고 판단하게 된다

이 방법을 이용하여 추출된 상,하위 관계에 있는 새로운 쌍의 예는 표 4와 같다.

상위어	하위어
Europe	UK
Computer Firm	Texas instruments
Remote users	Portable pcs
Applications developer	Lotus development corp
Word processors	Microsoft word
Laser printer	HP laserjet III

[표 4] 상, 하위 관계에 있는 새로운 seed

4. 실험 및 결과

실험을 위해 사용된 자료는 Text Retrieval Conference(TREC)에서 제공한 Ziff 문서(1989년-1990년)로 Ziff 문서는 총 785개의 문서를 포함하고 있다. 자료의 크기는 약 800 MB이며 그 중 268,610개의 단어를 포함하고 있는 1,001,000개의 문장을 이용하여 실험하였다. Ziff 자료의 예는 아래와 같다. 이중 abstract tag안의 내용만이 실험에 사용되었다.

```
<DOC>
<DOCNO> ZF32-244-004 </DOCNO>
<DOCID>09 754 449</DOCID>
<JOURNAL>Computerworld Jan 14 1991 v25 n2 p1(2)</JOURNAL>
<TITLE>3Com cuts back net plans. (3Com Corp.) (abandoning network operating system business)</TITLE>
<AUTHOR>Keefe, Patricia; Nash, Jim.&M;</AUTHOR><TEXT>
<ABSTRACT>3Com Corp abruptly announces its intention to withdraw from the
LAN operating systems market and focus its efforts on multivendor connectivity products.
</ ABSTRACT></TEXT>
<DESCRIPT>
Company: American Telephone and Telegraph Co. (Communication systems)
Ticker: COM
Topic: Fiber optics
Data Communications
T3 Communications
Feature: illustration
Caption: Higher in fiber.</DESCRIPT>
```

[그림 3] Ziff data의 예

	Snowball	Generalized Pattern
새seed의 개수	39	203
Precision	30.76 %	35.84%

[표 5] 일반화된 패턴을 이용해 실험한 결과

3.2장의 표2의 seed를 초기 seed로 하여 실험을 수행한 결과는 표5와 같다. ·

실험 결과 제안한 방법을 이용했을 경우 더 많은 seed를 찾는다는 것을 알 수 있다. 그러나 정확도에 있어서는 큰 차이가 없었다. 또한 C-value와 NC-value를 통해 선정된 단어 중 형용사가 포함된 단어의 경우 일반화된 방법을 사용하면 형용사가 제거되기 때문에 실험결과에 좋지 않은 영향을 미친다는 것을 알 수 있었다.

5. 결론 및 향후 과제

본 연구에서는 다양한 패턴을 인식하기 위해 일반화된 패턴을 사용하는 방법을 제안하였다. 일반화된 패턴을 사용하였기 때문에 여러 가지 형태로 존재하는 컨텍스트를 인식하여 기존의 방법보다 더 많은 양의 새로운 seed를 찾아낼 수 있었다. 그러나 정확도에 있어서는 일반화된 방법을 사용하는 것과 그렇지 않은 것 사이에 큰 차이가 없었다. 그리고 모든 형용사와 부사를 제거하는 것 보다 일부 의미 있는 형용사는 남겨두는 것이 더 좋은 결과를 가져올 수 있을 것이라고 판단된다.

참고문헌

[1] M.A.Hearst, Automatic acquisition of hyponyms from large text corpora , in Proceedings of the 14th international Conference on Computational Linguistics, 1992

[2] D. Alan Cruse, Hyponymy and its variety, The Semantics of Relationships An Interdisciplinary Perspective, 3-21, 2002

[3] Philipp Cimiano, Lars Schmidt-Thieme, Aleksander Pivk, Steffen Staab, Learning Taxo-nomic Relations from Heterogeneous Evidence, in Proceedings of the ECAI2004 Ontol-ogy Learning and Population Workshop, 2004

[4] F.Ciravegna, Adaptive information extraction from text by rule induction and generaliza-tion, in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCA I2001), (2001)

[5] Eugene Agichtein and Luis Gravano. "Snowball: Extracting Relations from Large Plain-Text Collections", In Proceedings of the ACM International Conference on Digital Li-braries (DL'00), 2000.

[6] Hang Cui, Min-Yen Kan, Tat-Seng Chua, Unsupervised Learning of soft patterns for generating definition, in Proceedings of 13th International World Wide Web Conference, 2004

[7] Eric Brill, Brill Tagger http://www.ling.gu.se/~lager/Home/brilltagger_ui.html

[8] Katerina T. Frantzi, Sophia Ananiadou, and Junichi Tsujii, The C-value/NC-value Method of Automatic Recognition for Multi-word Terms, Research and Advanced Tech-nology for Digital Libraries: Second European Conference, ECDL'98, 1998