온톨로지 인스턴스 생성을 위한 인물의 직위 정보 자동 추출

박윤희, 이영화, 이상조
경북대학교 컴퓨터공학과
{yhpark, allyh, sjlee}@sejong.knu.ac.kr

Automatic People Position Information Extraction for IT-People Event Ontology

Yun-Hee Park, Young-Hwa Lee, Sang-Jo Lee
Department of Computational Engineering, Kyungpook National University

요 약

온톨로지란 프로그램과 인간이 지식을 공유하는데 도움을 주기 위해 사용된 개념적 명세서로써 지식을 정형화하는 방법을 제시하여 추론의 기반을 제공하여 준다. 온톨로지 구축에 관한 기존 연구들은 스키마의 모델링에 초점을 두었다. 그러나 자료가 대용량화 됨에 따라 인스턴스를 자동으로 추출하는 기술은 온톨로지 구축에 꼭 필요하다. 이에 본 논문에서는 문서에서 인물의 직위 정보를 자동으로 추출하였으며, 문서 내에서 인물 상호 참조 처리를 통해 인물의 축약 명칭을 복원하였다. 또한 구 단위에서의 실패한 변동된 직위 정보는 중심 술어를 대상으로 격틀 정보를 완성해 나감으로써 확장한 결과 정확률 97.6%를 얻었다.

1. 서 론

온톨로지란 프로그램과 인간이 지식을 공유하는데 도움을 주기 위해 사용된 개념적 명세서이다. 최근에는 웹 2.0을 비롯한 많은 분야에서 온톨로지의 필요성이 증가되고 있으나 그 필요성에 비하여 구축의 자동률이 높지 않아서 온톨로지 구축에 많은 비용이 소요되고 있다.

온톨로지를 구축하는데 필요한 선행의 연구에서는 수집된 정보의 구조화 및 연결에 관한 스키마의 모델링 문제에 초점을 두었기 때문에 다양한 지식 유형 인스턴스를 자동화하는 데에는 크게 비중을 두지 않았다. 그러나 온톨로지 구축의 효율성 및 실용성을 고려할 때에 인스턴스 생성의 자동화는 반드시 필요한 기술이다. 이는 인지적 관점에서는 지식 획득의 자동화 과정이며, 온톨로지적 관점에서 보면 온톨로지 확장의 자동화이다.

본 논문에서는 IT 인물 정보 온톨로지를 구분하는 데 가장 기본이 되는 인물의 소속 및 직위 정보를 자동으로 추출하는 방법을 제안한다. 또한 추출된 인물 직위 정보를 이용하여 문서에서 반복적으로 나타나는 인칭대명사를 인물명으로 복원하였다. 이 결과 동일인에 대해 인칭 대명사로 인한 중복 인스턴스의 생성 방지가 가능하다. 왜냐하면 인칭 대명사의 인물명 복원으로 인해 온톨로지에 있는 기존 인물 인스터스와의 비교가 가능해지기 때문이다.

인물의 직위 정보를 자동으로 추출하는 본 논문의 구성은 다음과 같다. 제 2장에서는 추출된 인물의 직위 정보를 활용할 IT-People Event Ontology에 대해서 설명하고 3장에서는 인물 상호 참조 처리 통해 구 단위와 술어 단위의 직위 정보 추출을 설명한다. 제 4장에서는 실험 결과와 평가를 통해 오류를 분석하고 마지막 5장에서는 결론 및 향후 과제를 이야기 한다.

2. IT-People Event Ontology

인물 정보를 온톨로지로 표현하기 위해서는 영속성과 실제 활용을 고려해야 한다. 온톨로지가 상황에 따라 활용의 영속성을 유지하기 위해서는 개념과 관계를 어떻게 정의하는지가 중요하다. 온톨로지를 구축하기 위해서는 표현하고자 하는 대상의 본질은 무엇이고, 어떻게 표현할 것인지에 대한 철학적인 물음을 요구하기도 한다. 미조구찌는 우리가 표현하고자 하는 정보의 변하지 않는 본질적인 속성들을 온톨로지의 개념과 관계 정보로 표현하도록 권고하고 있다.[1] 하지만, 인물 정보는 시간에 따라 변할 수 있기 때문에 이를 수용할 수 있는 온톨로지의 스키마를 구축하는 방법이 요구된다.

인물에 관한 정보는 크게 두 가지로 나누어 볼 수 있다. 하나는 시간에 따라 변하는 정보이고 다른 하나는

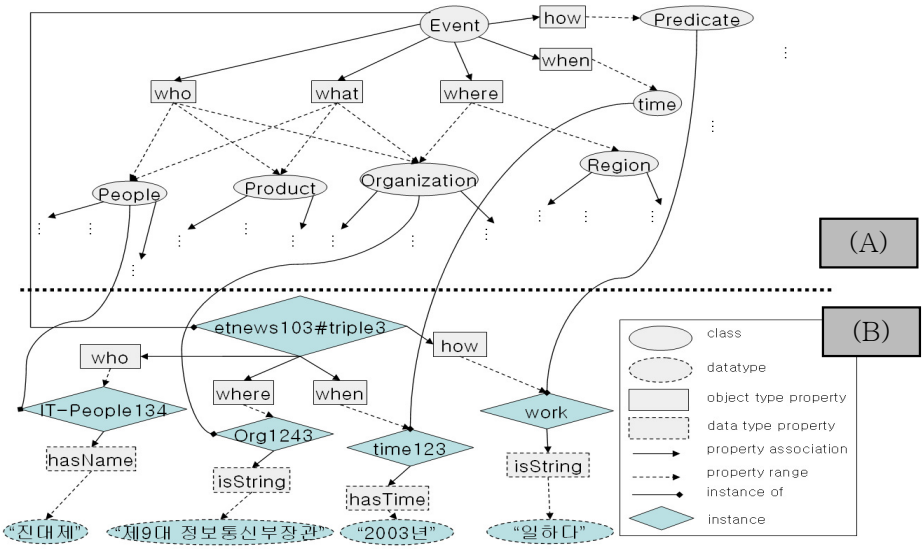

그림 1. IT-People Event Ontology

시간이 흘러도 잘 변하지 않는 정보이다. 예를 들면 인물의 이름, 성별, 생년월일, 출생지 정보는 잘 변하지 않는 정보에 속하지만 인물의 학력, 직업, 특히 직위는 잘 변하는 정보에 속한다. 만약 시간에 따라 변하는 정보들을 속성(Property)으로 정해둔다면 정보가 변할 때마다 속성(Property)을 수정해야 하므로 온톨로지의 궁극적인 본질에 맞지 않으며 인물 정보에 대한 역사성(History)도 없어서 이력을 찾기도 어려울 것이다. 이를 해결하기 위해 본 논문에서는 인물 정보가 가진 시간을 고려하여 인물에 관한 정보를 시간의 흐름에 따른 사건(Event)으로 간주하였다.

본 논문에서의 IT-People Event Ontology(이하 IT-PEO)[1]는 인물에 관해 값이 변하지 않는 속성(Property)과 값이 변하는 속성(Property)을 구분하여 담고 있다. 그림 1은 IT-PEO의 스키마와 하나의 인스턴스를 예를 들어 표현하고 있다. 그림에서 가운데 가로로 표현된 점선의 위쪽(A)은 온톨로지 스키마를 나타내고 있고, 아래쪽(B)은 온톨로지에 담겨 있는 정보 즉, 인스턴스의 예를 보여주고 있다.

3. 인물의 직위 정보 자동 추출

앞 장에서 설명한 IT-PEO는 웹 문서에서 IT 인물에 관한 정보를 추출하여 구축되고 있다. 그러나 IT의 인물과 관련된 기사에 담겨 있는 수많은 정보에서, 기반 기술의 부족으로 온톨로지 구축에 필요한 다양한 지식 유형을 추출 하는데 많은 어려움이 있다.

본 논문에서는 IT-PEO의 인스턴스 구축을 위한 인물 정보 자동 추출을 목적으로 개발되었다. 그림 2는 본 논문에서 제안하는 인물의 직위 정보 자동 추출 과정이다.

개체명이 인식된 웹 문서를 입력으로 받아들여 구 단위의 인물 직위 정보를 추출한다. 구 단위의 직위 정보 추출시 발생할 수 있는 중복 인스턴스 생성을 방지하기 위하여 인물의 직위 조합 정보 테이블을 이용하여 상호 참조를 처리한다. 구 단위에서의 인물 직위 정보 추출에 실패한 문장들은 술어 단위의 인물 직위를 추출로 확장해 나간다.

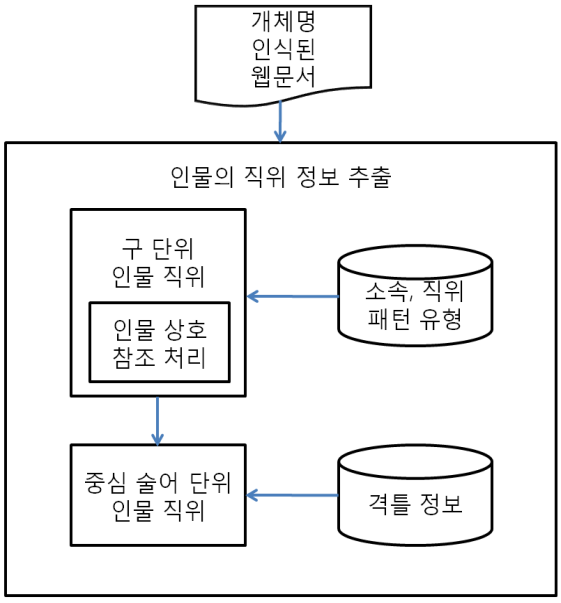

그림 2. 인물의 직위 정보 자동 추출 과정

1) IT-People Event Ontology는 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 구축되었습니다.

3.1 구 단위의 인물 직위 정보 추출

웹 문서에서 추출 가능한 인물에 관한 정보는 다양하나 현재로서는 문장에서 가능한 모든 인물의 의미적 정보를 자동으로 추출하는 것은 불가능하다. 본 논문에서는 전자 신문의 웹 문서를 대상으로 인물에 관한 소속 및 직위 정보를 자동 추출의 대상으로 한다. 본 논문이 대상으로 하는 인물의 직위 정보 추출은 인물의 정보가 비즈니스의 한 모델로 대두되고 있는 현대에 있어서 의미있는 작업이 될 것이다. IT-PEO의 인스턴스 대상은 작은 신생 기업인 또는 벤처 기업인도 다수라서, 이들의 정보를 수작업으로 입력하기에는 많은 비용이 발생한다. 따라서 자동화 방법을 이용하면 기존의 인물 정보 검색 시스템에서는 얻기 힘든 벤처 기업이나 신생 기업, 대중에게 잘 알려지지 않은 중견 업체의 인물 정보까지 획득할 수 있다는 장점이 있다.

또한 이러한 인물의 직위 정보의 자동 추출은 온톨로지의 초기 인스턴스 생성 단계 뿐만 아니라 기존 온톨로지에 인스턴스가 추가될 때에도 의미 있는 작업이다. 왜냐하면 동일명이 IT-PEO에 여러 명 존재할 경우에 그 인물들의 소속이나 직위 정보를 이용하여 동일 인스턴스인지를 판단하게 되는 중요한 근거가 될 수 있기 때문이다. 이 절에서는 전자 신문의 302개의 웹 문서의 4,463문장을 분석하여 53개의 패턴을 추출하였으며 인물의 소속과 직위에 관련된 정보의 패턴 유형 53개를 크게 분류하면 표 1과 같다.

표 1. 인물 직위 정보 관련 문장 패턴 유형

유형 1 : 단일 직위인 경우	
1	조직명+사람이름+직위명
2	조직명+직위명+사람이름
3	사람이름+조직명+직위명
4	조직명1+사람이름+직위명+조직명2
5	조직명1+“이하”+조직명2+사람이름+직위명
6	조직명1+“이하”+조직명2+직위명+사람이름
7	조직명1+직위명+사람이름+“이하”+조직명2
유형 2 : 공동 직위인 경우	
8	조직명+사람이름1+사람이름2+직위명
9	조직명+직위명+사람이름1+사람이름2
유형 3 : 지역명을 포함한 조직명인 경우	
10	지역명+조직명+사람이름+직위명
11	지역명+조직명+직위명+사람이름
12	조직명+지역명+직위명+사람이름
13	조직명+지역명+사람이름+직위명
유형 4 : 조직명이 변경된 경우	
14	조직명1+“구”+조직명2+사람이름+직위명
15	조직명1+“옛”+조직명2+직위명+사람이름
유형 5 : 겸임 직위인 경우	
16	조직명1+직위명+“겸”+ 조직명2+직위명+사람이름
17	사람이름+조직명1+직위명+“겸”+ 조직명2+직위명
18	조직명+지역명+직위명1+사람이름+직위명2

2006년 9월부터 2007년 9월까지 1년 동안의 전자신문의 직위와 관련된 웹 문서를 분석한 결과, 다음과 같은 결정 규칙을 얻었다.

▶ 결정 규칙 1
사람이름을 중심으로 여러개의 조직명[2]이나 직위명이 나올 경우 사람이름과 가까운 거리의 조직명과 직위명을 추출한다.

▶ 결정 규칙 2
지역명을 포함한 직위의 경우 ‘지역명 + 조직명’, ‘조직명 + 지역명’ 모두를 단일 조직명으로 인식하도록 한다.
예) 한국 IBM
다우 코리아

구 단위의 인물 직위 정보 추출을 위한 53개의 패턴을 이용 하였을 경우 대부분의 웹 문서에서 인물의 소속과 직위에 관련된 문장에서 이름과 가장 가까운 거리의 소속이나 직위 정보가 실제 인물의 소속 정보였다. 따라서 소속이나 직위 정보가 여러 개 나왔을 경우 사람이름과 거리상으로 가까운 소속, 직위 정보를 추출하도록 하였다. 그러나 이렇게 하였을 경우에는 인물의 이직에 관련된 직위 정보 추출에는 한계가 있다. 구 단위의 인물 정보 추출의 정확도 감소 원인은 바로 직위 변동과 관련된 술어 정보를 활용하지 않았기 때문이다. 이에 다음 절에서는 구 단위에서 추출에 실패한 직위 정보 추출을 위하여 술어 정보를 활용한다. 술어 정보를 활용 하였을 경우 인물의 이직이나 승진에 따른 직위의 변동 정보 추출이 가능하다.

3.2 한 문서 내에서 인물 상호 참조(Coreference) 처리

신문 기사에서 인물을 지칭하는 형태는 크게 축약형 인물 명칭과 원형 인물 명칭으로 나누어 볼 수 있다. 축약형 인물 명칭이란 한 문장에서 사용된 인칭 대명사가 이전 문장에 한번이상 나타난 인물을 가리킴으로써 축약된 형태를 가지는 경우를 말하며 원형 인물 명칭이란

2) 경제, 교육, 군사, 미디어, 스포츠, 예술, 공연, 전시, 미술관, 박물관, 사회기관, 의학, 종교, 과학, 통신, 교통, 기업, 도서관, 법률, 행정, 정치 관련 기관 단체와 회담, 회의 모두 포함

한 문장에서 새롭게 나타나는 인물 명칭으로 인물의 완전한 이름으로 표현된다. 다음은 인칭에 관련한 축약형과 원형 인물 명칭의 예를 보여주고 있다. 본 논문에서는 축약형 인물 명칭의 원형 복원 과정을 인물 상호 참조로 정의하고 이들간의 상호 참조를 해결한다.

◆**원형 명칭**

- **김○○씨**(a)는 30일 ....
- **그**(b)는 다음달 ....

◆**축약 인물 명칭**

- **김○○**(c) 과학기술부 장관은 ....
- **김**(d) 장관은 또 이△△ 우주항공관장과 ..

<축약 인물 명칭과 원형 명칭의 예>

이처럼 축약형 인물 명칭의 원형을 복원하는 이유는 축약형 인물 명칭은 온톨로지에서 기존의 인스턴스와의 중복을 발생시킬 수 있기 때문이다. 왜냐하면 개체명 인식 과정을 통한 개체명 인식 결과가 (a)와 (b), (c)와 (d) 모두 별개의 인물로 인식될 경우 (a)와 (b), (c)와 (d)가 동일 인물인지를 파악해야만 온톨로지의 기존 인물 정보를 추가할 것인지 아니면 새로운 인스턴스를 생성할 것인지를 결정할 수 있기 때문이다. 그림 3은 본 논문에서 제안하는 인물 상호 참조 처리 과정을 보여준다.

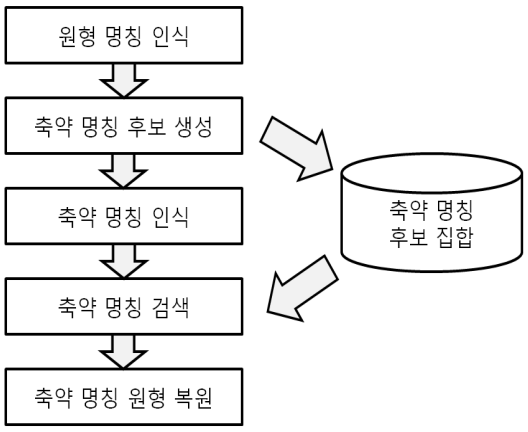

그림 3. 인물의 상호 참조 처리 과정

인물의 상호 참조 처리 과정에서는 일단 개체명 인식된 웹 문서로부터 원형 명칭을 인식한다. 인식된 원형 명칭을 바탕으로 생성 가능한 모든 축약형 인물 명칭의 후보 집합을 생성하고 이후의 문장에서 축약형 인물 명칭을 인식하게 되면 미리 만들어 놓은 축약 명칭 후보 집합을 검색하여 축약형 인물 명칭의 원형 명칭을 복원한다.

3.3 중심 술어 단위의 인물 직위 정보 추출

구 단위에서의 한계점으로 지적된 인물의 직위 변동에 관한 정보 추출을 위하여 술어 정보를 활용한다. 이때에 인물의 직위 정보 변동과 관련된 술어들을 중심 술어라고 정의하고 그 중심 술어들을 대상으로 직위 정보를 추출한다. 중심 술어는 아래와 같다.

▶ 중심 술어 유형(일부)
근무하다.
입사하다.
일하다.
선임하다.
발령하다.
영입하다.
발탁하다.
근무하다.
승진하다.
취임하다.
임명하다.

본 절은 이와 같은 중심 술어를 바탕으로 중심 술어가 필요로 하는 격틀 정보를 채워나간다. 먼저 격틀 정보를 채워나가기 위해서는 한국어의 문법 관계 분석이 필요하다. 그리고 한국어의 문법 관계 분석에는 한국어의 교착어적 특성상 조사가 중요한 단서를 제공한다. 대부분은 명사구의 조사로 그 문법 관계를 파악할 수 있는데 예를 들면 주격은 '-이/-가' 조사를 동반하고 목적격은 '-을/-를' 조사를 동반한다. 그러나 한국어의 언어 운용상 조사를 생략하거나 '-은/-는, -만, -도, -부터' 등의 보조사는 대부분의 문법 관계에 사용될 수 있기 때문에 문법 관계 분석시에는 이를 고려하여야 한다.

본 절에서는 술어를 중심으로 절 단위의 주어, 목적어, 부사어 중 하나의 문법 관계만을 고려하여 각각의 문법 관계를 독립적으로 분석한다. 이렇게 분석한 각각의 문법 관계를 이용하여 중심 술어가 필요로 하는 격틀 정보를 완성한다. 만약 우리가 도메인과 관련된 술어와 그 술어의 인수 중 어느 것이 중심 술어의 격틀 정보를 채우고 있는지의 두 가지 사항만을 알고 있다면

원하는 정보의 추출을 할 수 있다. 그림 4는 인물의 변동된 직위 정보 추출을 위해 정의된 격틀 정보와 그 격틀에 들어갈 내용들을 보여준다.

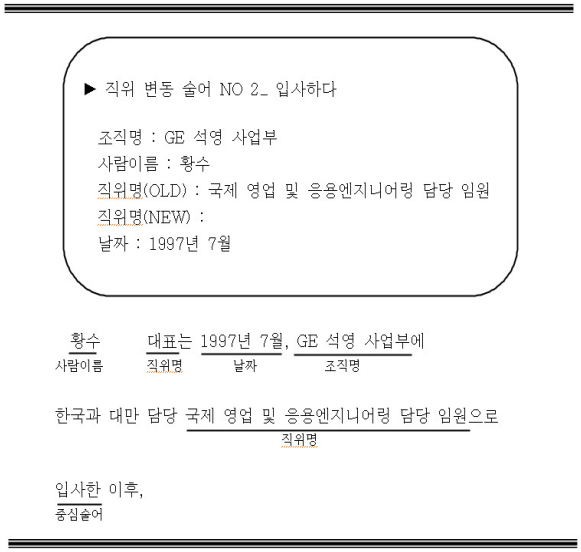

그림 4. 중심술어 “입사하다”의 격틀과 격틀에 들어갈 내용

중심 술어를 대상으로 직위 정보 추출시 날짜 데이터를 이용하여 직위명(OLD)와 직위명(NEW)를 판단한다. 날짜 데이터가 언급되지 않았을 경우에는 신문 기사의 날짜를 그 판단의 기준으로 삼는다. 그림 5는 또 다른 중심 술어인 “임명하다”의 격틀 정보와 그 내용을 보여준다.

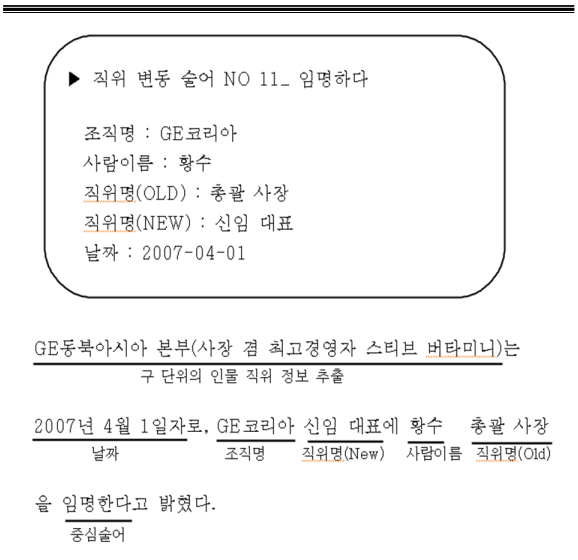

그림 5. 중심술어 “임명하다”의 격틀과 격틀에 들어갈 내용

5. 실험 및 평가

본 논문에서 제시하는 결과들은 2006년 9월부터 2007년 9월까지 전자신문 기사 중 512개의 문서를 대상으로 얻어진 결과들이다. 실험에는 86,012 어절의 4,463 문장을 사용했다. 실제 데이터를 분석하여 얻은 구 단위의 인물 직위 패턴을 적용하여 간단한 기본직위 정보를 추출하고 구 단위의 인물 직위 정보 추출에 실패한 문장을 모아 중심술어를 대상으로 절 단위의 인물 직위 정보를 추출하였다.

실험에 사용되지 않은 59,810어절의 3103문장을 평가에 사용하였고 정확률과 재현율로 평가하였다. 표 2는 제안된 방법의 평가 데이터에서의 성능을 나타낸다.

표 2. 평가 데이터 성능 분석 결과

	인물의 직위 정보 추출 (210개 기사문 평가)	
	구 단위	중심 술어 단위
정확률	98.9	96.3
재현율	97.6	95.4

평가 데이터에서의 오류로는 한 문장에서 중심술어, 조직명, 직위명이 모두 출현하였음에도 불구하고 사람이름이 출현하지 않아서 정보를 추출하지 않는 경우이다. 사람이름의 부재로 인해 발생하는 정보 추출의 실패를 막기 위해서는 차후에 생략된 주어를 복원하여 보완하여야 할 것이다. 그림 6은 주어 생략으로 인한 정보 추출 실패의 예를 보여준다.

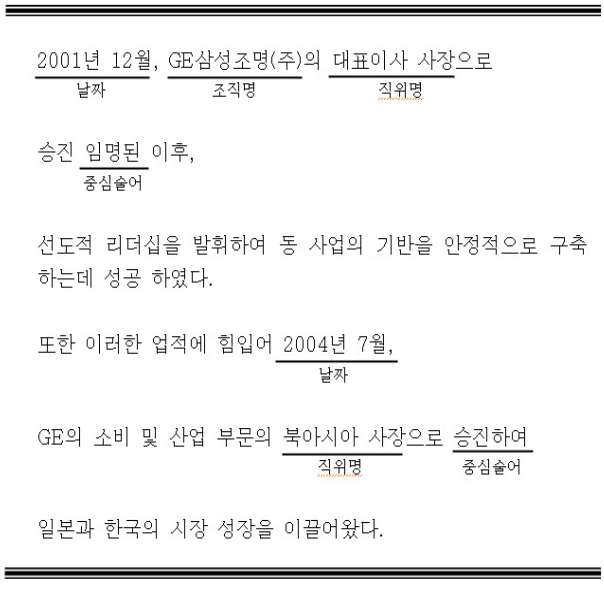

그림 6. 오류 유형의 예

6. 결론 및 향후 과제

본 논문에서는 텍스트로부터 추출 가능한 여러 가지 의미적 관계들 중에서도 IT-PEO를 위한 인물의 직위 정보만을 그 추출 대상으로 실험하였다. 개체명 인식이 끝난 하나의 웹 문서를 받아들여 구 단위의 인물 직위 정보 추출을 위한 53개의 패턴을 적용한다. 구 단위의 인물 직위 정보 추출이 실패한 문장들은 다시 모아 중심술어를 대상으로 격틀 정보를 채워나간다. 총 512개의 문서를 대상으로 실험에는 302개의 문서가 사용되었고 평가에 210개의 문서가 사용되었다. 직위 정보 추출 실험의 정확률은 97.6%였다.

향후 더 나은 성능과 신뢰할 만한 결과를 위해서 좀 더 많은 데이터를 보강하고 문장에서 주어의 생략으로 인해 추출이 어려워진 직위 정보를 찾아내기 위한 주어 복원 문제에 대한 연구를 진행해 나갈 계획이다. 또한 현재 제안된 방법을 다른 술어를 중심으로 확장할 수 있는 방법에 대한 연구도 계속해 나갈 것이다.

참고 문헌

[1] Riichiro M, “Tutorial on ontological engineering – Part 3: Advanced course of ontological engineering New Generation Computing,“ *OhmSha&Springer*, no. 2, vol.22, 2004

[2] Kim, J., and Moldovan, D., "Acquisition of linguistic patterns for knowledge-based information extraction", *IEEE Transactiops on Knowledge and Data Engineering,* no. 7, vol 5, p. 713-724, 1995.

[3] Soderland, S., "Learning information extraction rules for semi-structured and free text", *To appear in the Journal of Machine Learning*, 1998

[4] T. Berners-Lee, J. Hendler and O. Lassila, “The Semantic Web.” Feature Article, Scientific American no. 5, vol. 1, 2001.

[5] A. Okumura, T. Ikeda, and K. Muraki, "Text summarization based on information extraction and categorization using 5W1H," Journal of NLP 6, p. 27-44. 1999.

[6] J. Cao, D. Roussinov, J. Antonio Robles-Flores and J. F. Nunamaker Jr, "Automated Question Answering From Lecture Videos : NLP vs. Pattern Matching," *Proceedings of the 38th Hawaii International Conference on System Sciences*, Volume 01, p. 43- 50, 2005.

[7] Soo-Min Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," In *Proceedings of COLING-06,* 2006.

[8] Choi, Y. Cardie, C. Riloff, E., and Patwardhan, S, "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns," *Proceedings of HLT/EMNLP-05*, 2005.

[9] Gildea, D. and Jurafsky, D. "Automatic Labeling of semantic roles", *Computational Linguistics,* no. 28, vol. 3, p. 245-288, 2002.

[10] Soo-Min Kim and E. Hovy, "Determining the Sentiment of Opinions," In *Proceedings of COLING-04,* p. 1367-1373, 2004.

[11] 윤재민, 정유진, 이종혁 육하원칙 활성화도를 이용한 신문기사 자동 추출 요약, *정보과학회논문지*, 제31권, 제4호, p. 505-515, 2004.

[12] 경북대학교, 온톨로지 검증 및 온톨로지 기반 인스턴스 생성에 관한 연구, 최종 보고서, 2006.

[13] T. Gruber, "What is an Ontology?," http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

[14] 이행원, 취재보도의 실제, 나남출판, 1999.

[15] 김지영, 현장신문론, 도서출판 쟁기, 1996.