정보추출을 위한 고유명사 및 대용어 태깅

장성호*, 강승식*, 우종우*, 윤보현**
*국민대학교 컴퓨터학부, 첨단정보기술연구센터
**한국전자통신연구원 언어공학부
{mpquake, sskang, cwwoo}@cs.kookmin.ac.kr, ybh@etri.re.kr

Named Entity and Coreference Tagging for Information Extraction

Sung-Ho Jang*, Seung-Shik Kang*, Chong-Woo Woo*, Bo-Hyun Yun**
*School of Computer Science, Kookmin University and
Advanced Information Technology Research Center
**Linguistic Engineering Department, ETRI

요약

최근 정보추출에 대한 중요성이 점차 증가하면서 정보추출에서 필요로 하는 Named Entity와 Coreference, Information Extraction, Information Retrieval의 소개와 한국어에 대해 적용시키기 위한 정의와 방법을 제시한다. 또한, 대량의 문서에 대한 태깅을 효율적으로 수행할 수 있도록 Named Entity와 Coreference 태깅을 쉽게 할 수 있는 NE-CO 태깅 도구를 개발하였다. 이 태깅 도구를 이용하여 시험적으로 경제, 공연, 여행 분야의 300문서에 대한 말뭉치를 구축하였으며, 이 말뭉치는 한국어 정보추출 시스템을 개발하는데 기초 자료로서 활용될 예정이다.

1. 서론

영어권 국가에서는 1980년대부터 정보추출에 대한 학술회의를 개최하여 정보추출에 대한 높은 관심을 보이고 있다. 이러한 학술회의로는 MUC (Message Understanding Conference)와 TREC (Text Retrieval Conference)가 있으며, 최근 일본에서도 IREX(Information Retrieval and Extraction Exercise)가 개최되었다[1,2,6,7].

정보추출은 기존의 색인어 위주의 정보검색으로는 검색될 수 없었던 중요한 키워드를 좀 더 정확한 방법으로 추출하는 것으로서, 활용분야가 일반 문서 검색 시스템이나 웹 문서 검색시스템 등 정보 검색 시스템에서 넓게 활용될 수 있다[3,4,5].

정보추출 중에서 중요한 부분이 Named Entity (NE), Coreference(CO), Information Extraction(IE), Information Retrieval(IR)이다[8,9,10]. 본 논문에서는 MUC와 IREX에 대한 소개 및 한국어 문서들을 대상으로 한 정보추출 시스템을 구축하는데 필요한 자료를 구축하는 방법을 제시하고자 한다.

2. Named Entity

NE(Named Entity)는 이름이나 위치, 혹은 수식 표현 같은 것 중 유일성을 가지고 있는 것들을 추출해내는 것으로 고유 명칭(entity names), 시간 표현(temporal expressions), 숫자 표현(number expressions) 3가지로 구성된다. 추출된 표현은 entity(조직, 사람, 위치), times(날짜, 시간), quantities(돈, 백분율)를 확인할 수 있는 유일한 표현이 되어야만 한다. 그 외에도 전화번호나 주소같이 다른 표현도 추출될 수 있으나 "어떤 표현(단어)를 추출한다"보다 추출된 표현(단어)이 유일한 표현이어야만 한다는 것이 중요하다.

NE는 SGML 태그 형식을 사용하지만, SGML이 필요로 하는 TXT, HL, DATELINE, DD 같은 기본적인 태그는 사용하지 않으며, 단지 SGML 태그 중 ENAMEX, TIMEX, NUMEX을 사용하며, 각각 entity, times, quantities 표현에 대응된다. ENAMEX는 ORGANIZATION(회사, 정부조직 혹은 다른 조직적인 개체 이름), PERSON(사람이나 가족 이름), LOCATION(행정적이거나 지리적으로 정의된 위치의 이름)의 TYPE을 가지고, TIMEX는 DATE (날짜 표현), TIME(시간 표현)의 TYPE을 가지고, NUMEX는 MONEY(돈의 표현), PERCENT(백분율)의 TYPE을 가진다.

Named Entity의 평가를 위해서 결과는 SGML 형식이 되야 하며, 공백이나, 줄 바꿈은 사용될 수

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

없다. 그렇지 않으면, 출발 시점이 변할 수 있으며, 평가에 영향을 줄 수 있다.[2)]

Named Entity 태깅의 기본 형식은 다음과 같다.

<ELEMENT-NAME ATTR-NAME="ATTR-VALUE"> text-string</ELEMENT-NAME>

즉, HTML 태그와 같이 정의된 태그 사이에 태깅하려는 단어가 들어가면 된다. 예를 들면, "(주)삼성전자"는 다음과 같이 태깅을 한다.

<ENAMEX TYPE="ORGANIZATION">(주)삼성전자</ENAMEX>

태깅 형식에 대해 주의해야 할 사항을 살펴보면, "남한과 북한 사람들"과 "남한 사람들과 북한 사람들"같이 생략된 표현이 있는 경우는 생략되지 않은 표현을 태깅할 때와 같이 해야 한다는 것이다. 즉, 다음과 같이 "남한과 북한"을 하나로 태깅하지 않고 "남한", "북한"을 따로 태깅을 한다.

<ENAMEX TYPE="LOCATION">남한</ENAMEX>과<ENAMEX TYPE="LOCATION">북한</ENAMEX> 사람들

<ENAMEX TYPE="LOCATION">남한</ENAMEX> 사람들과 <ENAMEX TYPE="LOCATION">북한</ENAMEX> 사람들

그러나 "한글과 컴퓨터 주식회사"같이 생략된 표현이 있는 것처럼 보이는 문장("한글과 컴퓨터"를 기업 이름이라고 할 때)이라도 생략된 표현이 없는 경우는 하나의 단위로 태깅을 해야 한다.

<ENAMEX TYPE="ORGANIZATION">한글과 컴퓨터</ENAMEX> 주식회사

이처럼 태깅해야 하는 부분이 애매한 경우가 발생할 수 있으며, 이러한 애매한 경우를 최대한 줄이기 위해서 한국어에 맞게 새로운 Named Entity 태깅 규칙을 작성해야 한다.[3)]

3. Coreference

Coreference는 관련성 있는 것들을 추출해서 그것들 사이의 관계를 링크하는 것이며, 태깅된 것들 사이에서 특별한 관계를 링크한다. 그래서 태깅된 링크는 선행된 다른 명확하고 관련성 있는 태그로부터 추론될 수 있으며, 이러한 태깅된 링크들이 체인 형식으로 구성될 수 있다.(명사구 A가 B와 관련이 있고, B가 C와 관련이 있다면, C 또한 A와 동일한 개체를 가리키게 된다.)

Coreference 역시 SGML 태그를 사용하며, 사용되는 SGML 태그의 속성으로는 TYPE, ID, REF, MIN, STAT가 있으며, TYPE속성은 앞에서 태깅되었던 것과의 관계를 가리키는데 사용되며, ID와 REF속성은 두 문장 사이의 관련된 링크를 가리키게 된다. MIN속성은 COREF 태그가 사용될 때 태깅된 내용의 최소한의 의미를 값으로 사용하며, STAT는 COREF로 태깅된 두 문장 사이의 관계가 애매모호한 관계일 경우 선택적으로 사용되며, 값으로는 'OPT'가 사용된다.

태깅를 해야 하는 표현은 명사, 명사구, 대명사(인칭대명사와 지시대명사)와 Named Entity에서 태깅된 것들에 대해서 할 수 있다. 그러나, Named Entity에서 태깅 되었다고 해서 모두 해야 하는 것은 아니며, Named Entity에서 태깅된 것의 substring은 태깅하지 않고, 명사화 접미사가 붙은 동사(영어의 동명사) 역시 태깅하지 않는다.

Coreference의 태깅 형식은 다음과 같다.

"길동은 사랑에 빠졌다. 그가 사랑하는 ...."

<COREF ID="100">길동</COREF>은 사랑에 빠졌다. <COREF ID="101" TYPE="IDENT" REF="100">그</COREF>가 사랑하는 ...

Coreference는 Named Entity와 함께 정보추출에서 상당히 중요한 부분을 차지한다. 만약 위의 예에서처럼 Named Entity에서 PERSON 개체로 "길동"을 찾아냈고, Coreference에서 "그"를 "길동"과 동일개체로 추출하였다면, 기존의 색인어의 출현 빈도만으로 용어 가중치를 구하는 정보 검색 시스템에서는 PERSON 개체(위의 예제에서는 "길동")의 출현 빈도가 한 번이므로 그 개체의 문서에 대한 관련성은 낮다.(실제로는 "길동"의 관련성이 높다.) 그러나 Coreference에 의해 추출된 PERSON 개체를 이용한다면, 기존 정보검색 시스템에서 낮았던 "길동"의 중요도가 높아지게 된다. 따라서 Coreference는 기존의 단순 출현 빈도만으로 계산되던 용어 가중치를 좀 더 향상시킬 수 있는 방법이며, Named Entity를 선행해서 Coreference를 수행한다면, 더욱 더 좋은 결과를 얻을 수 있다.

4. Information Extraction

Information Extraction(IE)는 Named Entity와 Coreference에서 추출된 정보를 특정한 structure (template)의 형식에 맞게 채워 넣은 것이다. IE는 두 개의 하위 작업으로 이루어지며, "Scenario Template"과 "Template Element"가 있다. "Scenario Template"은 항공 사고나 노사 협상 같은

2) Named Entity에 대한 평가는 "MUC-6 Scoring System User' s Manual"을 참조하면 자세한 정보를 얻을 수 있다.

3) MUC-6의 "Tokenization Rules"과 "Named Entity task Definition"의 부록을 참조하면 영어에 대한 정확한 태깅 규칙을 알 수 있다.

특정 시나리오에 종속적이며, 그 평가 방법으로는 template이 정확하게 instantiated objects를 포함하고 있는지와 시나리오 정의에 명확하게 slot (template의 member)을 채웠는지를 평가한다. "Template Element"는 "Scenario Template"과는 다르게 특정 시나리오에 독립적이며, 그 개체로 인명(person), 기관명(organization), 제품명(artifact)을 가진다. 이것들은 Named Entity나 Coreference 다음 단계에서 수행되어야 한다.

```
<PERSON> :=
        PER_NAME:     "NAME"^
        PER_ALIAS:    "ALIAS"*
        PER_TITLE:    "TITLE"*
        OBJ_STATUS:   {OPTIONAL}-
        COMMENT:      " "-
```

그림 1. Template Element의 PERSON object 예

그림 1을 보면, NE와 Coreference에서 추출된 정보가 Object와 slot에 채워진다. 각 object와 slot에는 특정한 규칙에 의해서 값이 채워질 수 있다. 예를 들면, <PERSON>의 PER_ALIAS의 경우 PER_NAME의 다른 명칭이 쓰여질 수 있으며, 당연하겠지만 PER_NAME에 값이 할당되지 않을 경우에는 PER_ALIAS는 의미가 없으므로 값을 넣은 것 자체가 무의미하다. 이 안에 채워지는 값의 정확성에 의해서 IE의 평가가 이루어진다. 결국 Named Entity와 Coreference에서 추출된 정보를 이용하여, 정보추출에 필요한 정보를 얻는 것이다.

5. Information Retrieval(IR)

IR은 문서의 집합으로부터 주어진 토픽에 관련된 문서를 추출하는 일이다. 일본에서 개최되었던 IREX(Information Retrieval and Extraction Exercise)에서는 문서의 집합으로 신문기사 2년 치의 분량을 사용하였고, 토픽은 다음의 예제처럼 토픽에 대한 ID와 DESCRIPTION, NARRATIVE로 구성된다.

```
<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>Corporate merging
</DESCRIPTION>
<NARRATIVE> The article describes a
corporate merging and in the article, the name
of the companies have to be identifiable.
Information including the field and the purpose
of the merging have to be identifiable. Corporate
merging includes corporate acquisition, corporate
unifications and corporate buying.
</NARRATIVE>
</TOPIC>
```

그림 2. IR 토픽에 대한 예제

IREX에서 참가자들이 제출한 Answer set에 대한 심사는 먼저 두 명의 사람이 심사를 하게 되고, 두 사람 서로 다른 심사를 하게 되면, 참가자가 없는 그룹의 지원자가 최종 심사를 하게 된다.

정답에 대한 평가는 3부분으로 이루어지며, 첫 번째는 원문의 주제에 대해 토픽이 일치하는 부분(A)과 토픽과 원문의 주제와는 일치하지는 않지만, 원문의 일부가 토픽과 관련이 있는 부분(B), 그리고, 원문의 주제와는 일치하지 않는 부분(C)으로 이루어진다. 이 세 부분을 이용하여 IR에 대한 평가가 이루어지게 된다.

6. NE-CO 태깅 도구

NE-CO 태깅 도구는 대량의 문서에 대하여 Named Entity와 Coreference 태깅을 효율적으로 수행할 수 있도록 도와 준다. 태깅 화면의 예를 그림 3과 같다.

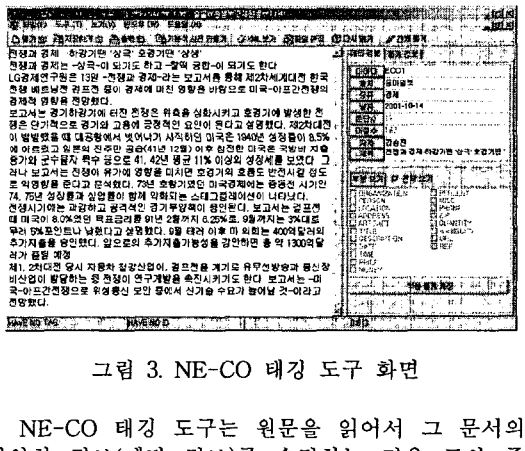

그림 3. NE-CO 태깅 도구 화면

NE-CO 태깅 도구는 원문을 읽어서 그 문서의 다양한 정보(메타 정보)를 수정하는 것을 도와 줄 수도 있으며, 또한 태깅을 하는 문서뿐만 아니라 이미 태깅이 되어 있는 문서들에 대해 태깅에 대한 다양한 통계를 보여줌으로써 태그별 태깅 정보를 쉽게 얻을 수 있게 만든 도구다.

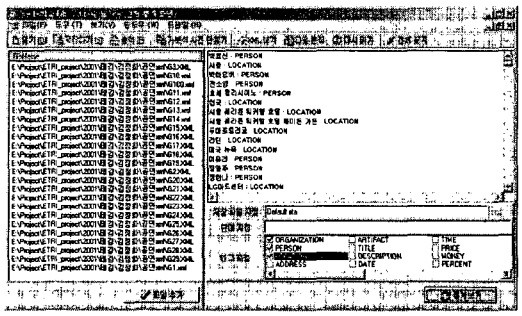

그림 4. NE-CO 태깅 도구 통계 화면

그림 4는 NE-CO 태깅 도구에서 통계를 내는 화

면으로 여러 개의 태깅된 문서를 선택해서 지정된 태그가 태깅된 단어를 보여준다.

이 태깅 도구를 이용하여 시험적으로 경제, 공연, 여행의 범주에 속하는 300개의 한글 문서에 대한 태깅을 수행하였다. 경제와 공연은 신문기사를 원칙으로 하고 문서당 400자 정도의 분량을 기준으로 한다. 여행 분야는 웹상에서 여행 관련 문서(가급적 여행지 소개, 교통편, 숙박, 입장료 등이 한 문서에 포함된 것)를 수집하여 구성한다. 각 분야별로 100개의 기사를 수집하여 태깅한다. 문서의 메타 정보에 문서 작성자, 문서 출처(웹문서일 경우, 웹페이지 주소), 문서의 타입(경제/공연/여행), 문서 작성일, 문장 수 등을 기술한다. 태깅된 문서 중 210개는 학습용으로, 90 문서는 평가용으로 활용할 예정이다. 태깅 문서의 예는 그림 5와 같다.

```
<DOC>                                   // 문서의 시작을 표시
<META>                                  // 문서의 메타 정보 기술
<DOC_ID> A13 </DOC_ID>                  // 문서의 고유 번호
<DOC_SRC> 한겨레신문 </DOC_SRC>         // 문서 출처
<DOC_TYPE> 여행 </DOC_TYPE>             // 문서의 종류 : 경제/공연/여행
<DOC_DATE> 2001년 8월 28일 </DOC_DATE>          // 문서 작성일
<DOC_SEN> 5 </DOC_SEN>                  // 문서에 포함된 문장 수
<DOC_EOJEL> 123 </DOC_EOJEL>            // 문서에 포함된 어절 수
<DOC_AUTHOR>김수웅</DOC_AUTHOR>         // 문서 작성자
</META>                                 // 메타 정보의 끝
<TITLE> 출장 보고</TITLE>               // 문서의 제목
<TEXT>                                  // 문서 본문의 시작
<S><PERSON ID=A13-1>철수</PERSON>는 <DATE ID=A13-2>8월 27일</DATE>에
<LOCATION ID=A13-3>대구</LOCATION>에 갔다.</S> <S><REF ID=A13-1>그
</REF>는 <REF ID=A13-3>그 곳</REF>으로 출장을 가는 도중에 <LOCATION
ID=A13-4>경주</LOCATION>에 들러 관광을 했다.</S> <S><REF ID=A13-2>그
날</REF> <REF ID=A13-4>그 곳</REF>의 날씨는 무척 더웠다.</S>
<S><REF ID=A13-1>그</REF>는 <LOCATION ID=A13-5>불국사</LOCATION>를
관람했는데, 입장료가 <MONEY ID=A13-6>1,400원</MONEY>으로 무척 저렴하
다고 생각했다.</S> <S><REF ID=A13-6>그런 가격</REF>이면 <REF
ID=A13-5>그 곳</REF>을 유지하는 비용으로 적당할지 생각해 보았다.</S>
</TEXT>                                 // 문서 본문의 끝
</DOC>                                  // 문서의 끝
// <S>…</S> 태그는 한 문장의 시작과 끝을 표현
```

그림 5. 한글 문서의 NE-CO 태깅 예

7. 결론

본 논문에서는 MUC과 IREX의 태깅 방법을 고찰하고 한국어에 대한 적용 방향을 모색하였으며, 한국어 태깅 도구를 구축하였다. Named Entity, Coreference, Information Extraction, Information Retrieval은 각각 서로에 대해서 어느 정도의 연관 관계를 가진다.

한국어에 적용시키기 위해 각 방법에 대해서 한국어 특성에 맞는 규칙 등을 연구해야 한다. 대표적인 예로 Named Entity에서 ENAMEX 태그의 경우 같은 어절에서 문맥에 따라 LOCATION이나 ORGANIZATION의 TYPE이 다르게 적용되는 경우가 발생할 수도 있다. 이것은 문맥에 의해 적절한 태깅을 적용해야 하지만, 사람의 주관이 들어가는 경우가 발생할 수 있기 때문에 태깅 오류가 발생할 수 있다는 것을 의미한다. 그러므로 위에서 언급한 태깅 방법을 좀 더 명확하고 한국어의 특성에 적합한 세부 규칙들이 필요하다.

참고문헌

[1] MUC-6, Homepage http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[2] IREX, Homepage http://cs.nyu.edu/cs/projects/proteus/irex/index.html

[3] 정래준, 김준태, "고유 명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구", 한글 및 한국어 정보처리 학술발표논문집, pp 68-72, 1996.

[4] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류", 정보과학회 추계 학술발표회,pp 170-172, 2000.

[5] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출", 한글 및 한국어 정보처리 학술발표논문집, pp 11-18, 2000.

[6] MUC-6, Proceedings of 6th Message-Understanding Conference (MUC-6), Morgan Kaufmann, 1995.

[7] MUC-7, Proceedings of 7th Message Understanding Conference(MUC-7). MUC. 1998.

[8] Cardie, C., "Empirical Methods in Information Extraction", AAAI-97, pp.65-79, 1997.

[9] Appelt, D. E. and David J. Israel, "Introduction to Information Extraction Technology", A Tutorial Prepared for IJCAI-99, 1999.

[10] Riloff, E., "Information Extraction as a Stepping Stone toward Story Understanding", Understanding Language Understanding: Computational Models of Reading, MIT Press, 1999.

[11] Riloff. E., "Automatically Generating Extraction Patterns from Untagged Text", Proceedings of the thirteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.1044-1049, 1996.

[12] Sekine, S., and Y. Eriguchi, "Japanese Named Entity Extraction Evaluation - Analysis of Results", the 18th International Conference on Computational Linguistics (COLING'2000), pp.1106-1110, 2000.

[13] R. Mitkov, "An Integrated Model for Anaphora Resolution", Proceedings of the 15th International Conference on Computational Linguistics COLING'94, pp.1170-1176, 1994.

[14] Breck Baldwin, "CogNIAC: High Precision Co-Reference with Limited Knowledge and Linguistic Resources", ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, pp.38-45, 1997.

[15] Utsuro, T. and M. Sassano, "Minimally Supervised Japanese Named Entity Recognition Resources and Evaluation", In Proc. Of the 2nd International Conference on Language Resources and Evaluation, pp 1229-1236, 2000.