DOI : 10.5392/JKCA.2010.10.12.457

국방 기사 데이터를 이용한 맞춤형 정보 분석 시스템

Customized Information Analysis System Using National Defense News Data

최중환, 임채오
국방기술품질원

Jung-Whoan Choi(c0802@hanafos.com), Chea-O Lim(yoosep23@hanmail.net)

요약

맞춤형 정보 분석 시스템이란 정형화 되어 있지 않은 자연어 텍스트에서 유용한 정보를 추출하고 고객이 요구하는 맞춤형 정보로 가공하여, 미래를 예측하거나 추론하는데 도움을 주는 시스템을 말한다. 이러한 정보 분석 시스템을 구현하기 위해서는 자연어를 분석하는 자연어 처리 기술과 텍스트에서 필요한 개체와 그것들의 관계를 찾아내는 정보 추출 기술, 추출한 데이터로부터 알려지지 않은 새로운 정보를 찾아내는 데이터 마이닝 기술이 필요하다. 본 논문에서는 국방 기사 데이터를 대상으로 맞춤형 정보 분석을 수행하는 가상의 시스템을 제안하고, 정보 분석을 위한 기반 기술들을 소개한다.

■ 중심어 : | 정보분석 | 맞춤형 정보분석 시스템 | 국방기사 |

Abstract

Customized information analysis system is a software system that can help to extract useful information from non-structured natural language data, process the information to customized form, and provide future forecast and reasoning information. To implement the information analysis system, we need natural language processing technology to analyze natural language, information extraction technology to detect necessary entity and its relationship from text, and data mining technology to discover new and unknown information from extracting data. This paper suggest virtual customized information analysis system processing national defense news data and introduce base technologies for information analysis.

■ keyword : | Information Analysis | Customized Information Analysis System | National Defense News |

I. 서 론

오늘날에는 인터넷의 발달과 함께 뉴스기사, 블로그, 카페, 웹페이지 등 수많은 문서가 계속해서 생성되고 있다. 끊임없이 만들어지는 문서들 속에서 원하는 내용을 포함하고 있는 문서를 찾아내어 필요한 내용을 추출하고 그 데이터를 분석하여 유용한 정보로 재생산 하는 것은 많은 시간과 비용이 드는 작업이다. 특히 국방 분야의 정보 분석은 업무의 특성상 매우 높은 정확도를 요구하지만 제한된 인력으로 하루에도 수십 건씩 업데이트되는 문서들을 사람이 직접 분석하는 것은 매우 힘든 일이다. 사람이 모든 문서를 다 확인할 수 없기 때문에 현재의 정보 분석은 필요한 정보의 요구가 있을 때 그와 관련된 각 연도별 문서를 찾아서 수작업으로 통계

접수번호 : #101007-008
접수일자 : 2010년 10월 07일
심사완료일 : 2010년 11월 29일
교신저자 : 최중환, e-mail : c0802@hanafos.com

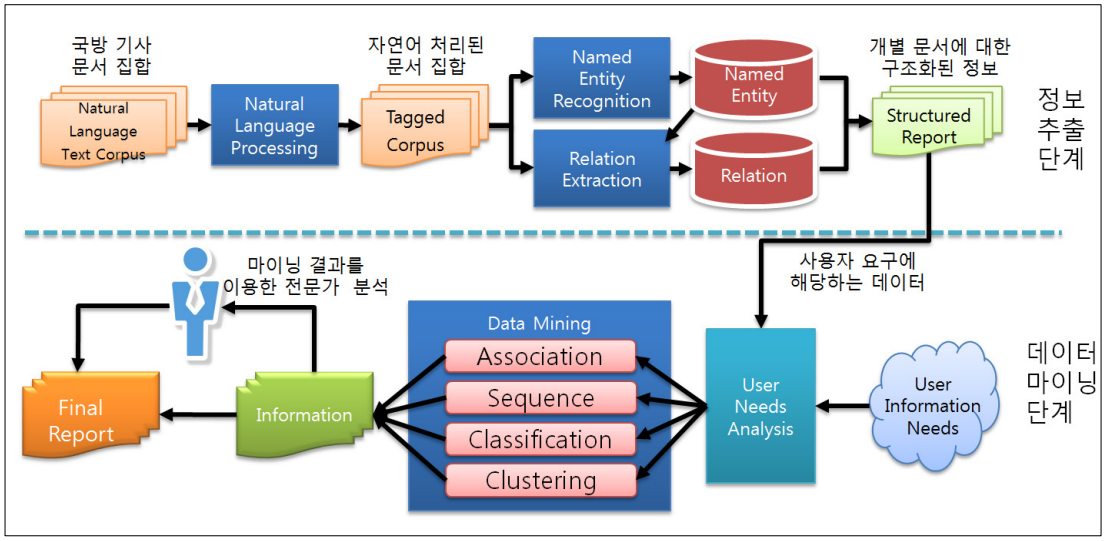

그림 1. 맞춤형 정보 분석 시스템 구성도

자료를 작성하고 그 정보를 이용하여 전문가가 직접 분석을 수행하고 있다. 하지만 이렇게 사람이 직접 분석하는 방법은 대량의 문서를 참고할 수 없어 신뢰성 있는 결과를 얻기 어렵고 많은 비용이 소요된다. 이러한 문제를 해결하기 위해 본 논문에서는 정형화 되어 있지 않은 자연어 텍스트로부터 자동으로 유용한 정보를 추출하여, 그 속에서 새로운 정보를 찾아내기 위한 기반 기술들을 소개하고, 국방 기사 데이터를 대상으로 정보 분석을 수행하는 가상의 시스템을 제안한다.

자연어로 된 문서들을 처리하기 위한 기반 기술들을 자연어 처리 기술이라고 하는데, 세부적으로 품사 부착, 구문구조 분석, 의미 분석 등이 있다. 이들 기술을 이용하여 분석한 문서로부터 필요한 정보를 추출하는 작업을 정보 추출이라고 하는데, 그 중에서 고유의 이름과 숫자 표현 등을 추출하는 것을 개체명 인식이라 하고, 개체명들 간의 관계를 알아내는 것을 관계 추출이라 한다. 정보 추출 기술을 이용하여 추출한 개체명과 그 관계를 데이터베이스화 하여 저장한 후, 그 데이터베이스 내에 존재하지만 숨겨져 있는 상관관계, 패턴, 추세 등의 새로운 정보를 발견하는 작업을 데이터 마이닝이라 한다.

본 논문은 Ⅱ장에서 맞춤형 정보 분석 시스템을 소개하며, 각 모듈별의 역할과 필요한 기술들에 대하여 알아본다. Ⅲ장에서는 실제 국방 기사 데이터를 제안하는 시스템으로 처리하는 예제를 보여주며, Ⅳ장에서 본 논문의 결론을 제시한다.

Ⅱ. 맞춤형 정보 분석 시스템

본 논문에서 제안하는 시스템([그림 1])은 현재 연구되어 있는 기술들을 이용하여 구조화 되어 있지 않은 국방 기사 텍스트 데이터로부터 정보를 추출하여 구조화된 정보(데이터베이스)로 변환하고, 그 데이터베이스로부터 사용자의 요구를 반영한 새로운 정보를 추출하는 맞춤형 정보 분석 시스템이다. 이 시스템은 크게 자연어 데이터로부터 필요한 개체명과 관계를 추출하여 데이터베이스화 하는 정보 추출 단계와, 이렇게 만들어진 데이터베이스와 사용자의 정보 요구를 분석하여 데이터베이스로부터 새로운 정보를 생성하는 데이터 마이닝 단계로 나누어진다. 이 장에서는 맞춤형 정보 분석을 위해 필요한 모듈과 각각의 역할, 필요 기술에 대하여 알아본다.

1. 자연어 처리 모듈(Natural Language Processing)

자연어는 사람이 인공적으로 만들어낸 인공어의 반의어로, 인간사회의 형성과 함께 자연발생적으로 생겨나고 세월의 흐름과 함께 진화하며 일상의 생활 속에서 서로 의사소통을 행하기 위한 수단으로서 사용되는 언어이다. 이러한 자연어는 많은 모호성을 가지고 있기 때문에 모호성을 해소하고 자연어를 전산적으로 이용할 수 있는 형태로 변환하는 자연어 처리 모듈이 필요하다. 자연어 처리는 크게 세가지 작업으로 이루어진다.

1.1 품사 부착

'paper'라는 단어를 보자. 'paper'는 문장에서 '종이'나 '서류'를 뜻하는 명사로 쓰일 수도 있지만, '종이의', '얇은'이란 형용사로 쓰일 수도 있고, '종이에 쓰다', '종이로 싸다'같은 동사로 쓰일 수도 있다. 품사 부착이란 이런 단어 단위의 모호성을 해소하기 위하여 텍스트 내의 단어들에 특정 품사를 붙이는 과정이다.

초기에는 주로 대상 단어와 앞 뒤 문맥을 고려한 규칙 기반의 연구가 많이 시행되었다. 하지만 언어는 신조어가 계속해서 생성되고, 새로운 언어가 생성될 때마다 새로운 규칙을 찾는 것은 계속해서 많은 비용이 소요되는 작업이다. 따라서 현재는 규칙 기반 방법보다 품사가 부착되어 있는 일정량의 말뭉치로부터 학습하여 모델을 생성하고, 그 모델을 이용하여 새로 입력되는 말뭉치에 품사를 부착하는 기계 학습 기반의 연구[1]가 주를 이룬다. 품사가 부착되어 있는 일정량의 말뭉치를 만드는 것 또한 적지 않은 비용이 드는 작업이기에 최근에는 품사가 부착된 학습 데이터 없이 일반 텍스트만을 사용하여 학습하는 자율 학습 기반의 연구[2]들이 활발히 진행되고 있다.

1.2 구문 구조 분석

구문 구조 분석이란 문법을 이용하여 문장을 구성하고 있는 구성 성분으로 분해하고, 그들 사이의 의존 관계를 분석하여 문장의 구조를 결정하는 작업이다. 다음의 문장을 보자. "아름다운 그녀의 목소리를 들었다."라는 문장은 2가지 의미로 해석될 수 있다. '그녀'가 아름답다는 의미일 수도 있고, '목소리'가 아름답다는 의미일 수도 있다. 구문 구조 분석은 문장의 구조를 밝히고 정확한 의미를 분석하는데 도움을 준다.

구문 구조 분석은 크게 구조적 파싱(Syntactic Parsing)과 의존 파싱(Dependency Parsing) 방법으로 나뉜다. 구조적 파싱은 일련의 문자열을 의미 있는 토큰(token)으로 분해하고 이들로 이루어진 파스 트리(parse tree)를 만드는 파싱 방법이다. 그림 2는 구조적 파싱 방법으로 "Last week IBM bought Lotus."이란 문장을 파싱한 예이며, 대표적인 연구로[3]이 있다.

의존 파싱은 단어 사이의 의존 관계들만을 밝히는 파싱 방법이며 대표적인 연구로는[4]가 있다. 잘 알려진 영문 파서인 Collins Parser[3], MaltParser[4], Stanford Parser[5] 등이 모두 공개되어 있기 때문에 영문 구문 구조 분석에 활용할 수 있다.

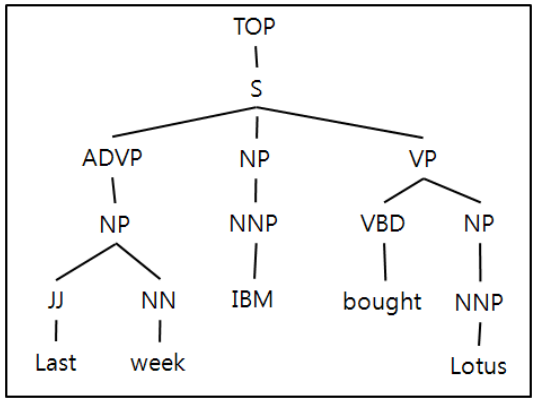

그림 2. 구조적 파싱 결과 예제

1.3 의미 분석

의미 분석이란 자연 언어의 의미를 파악하고, 이를 정규화된 형태로 표현하는 것이다. 의미 분석에는 의미역 부착(Semantic Role Labeling)과 단어 의미 중의성 해결(Word Sense Disambiguation)이라는 두 가지 분야가 있다.

의미역 부착이란 문장에서 술어의 의미 논항을 판별하고 각 논항의 의미역을 분류하는 것이다. 문장에서 언어의 구성 요소(단어, 구 등)들은 어떤 종류의 의미역

을 가지는데, 술어와 그 술어의 논항을 분석함으로써 문장의 대략적인 의미 구조를 알 수 있고, 여러가지 변형된 문법에서 동일한 의미의 문장을 인식할 수 있다. 다음 두 문장을 보자.

Tarzan kissed Jane.

Jane is kissed by Tarzan.

두 문장은 동일한 의미를 가진다. Tarzan이 어떤 행동(kiss)의 주체이고 Jane이 이 주체로부터 영향을 받는 대상이라는 것을 파악할 수 있다면, 위의 두 문장이 문법적으로는 다르지만 같은 의미라는 것을 파악할 수 있다.

단어 의미 중의성 해결은 주어진 문장에서 사용된 단어의 적합한 의미를 결정하는 문제이다.

A mother crane soon laid an egg.

'crane'은 두루미와 기중기라는 두 가지 의미를 가지고 있지만 우리는 이 문장에서 'crane'이 두루미라는 의미를 어렵지 않게 유추할 수 있다. 단어 의미 중의성 해결은 이런 동음이의어, 동형이의어, 다의어 등 여러 의미를 가지는 단어가 문장 내에서 가지는 의미를 찾아내는 작업이다. 가장 단순한 방법으로 같은 문장에서 함께 출현한 빈도를 계산하는 방법이 있다[12]. '두루미'라는 뜻의 'crane'은 'laid', 'egg' 등의 단어와는 매우 빈번히 함께 출현 하겠지만 '기중기'라는 뜻의 'crane'은 그렇지 않기 때문이다.

자연어 처리 모듈은 실제 국방 기사 텍스트 데이터를 입력받아서 시스템에서 사용할 수 있는 형식으로 변환(전처리)하고, 품사 부착, 구문 구조 분석 등을 수행하여 개체명 인식과 관계 추출 모듈에서 사용할 수 있는 데이터로 변환해 준다. 영문 자연어를 대상으로 한 품사 부착기와 구문 구조 분석기 등은 공개된 프로그램이 많이 있기 때문에 기존의 프로그램을 활용할 수 있다. 자연어 처리 모듈에서 얻은 정보(품사, 구문 구조, 공기 빈도)를 자질로 하여 개체명 인식과 관계 추출을 수행하게 된다.

2. 개체명 인식 모듈(Named Entity Recognition)

자연어로 된 문서 내에서 필요한 개체를 인식하는 것은 개체명 인식이라 하고, 인식된 개체명들 간의 관계를 추출하는 작업을 관계 추출라고 한다. 이 두 가지 단계를 합쳐서 정보 추출이라고 부른다.

개체명 인식 과정이 필요한 이유는 자연어 문서에서 국가, 조직, 무기명, 금액 등의 개체명을 추출하여 구조화된 데이터베이스로 만들기 위함이다. 개체명 인식 과정을 거쳐 구조화된 데이터로 변환해야 통계 자료를 뽑거나 데이터 마이닝을 거쳐 그 속에 숨겨진 정보를 찾아낼 수 있다.

개체명 인식에 관한 연구는 MUC(Message Understanding Conference)를 중심으로 이루어졌으며, 주로 사람, 단체, 지역명과 같은 고유의 이름들과 시간, 날짜, 액수, 퍼센트 표현과 같은 숫자 표현을 인식하는 것이다. [그림 3]은 "Jim bought 300 shares of Acme Corp. in 2006."라는 문장에 있는 사람, 양, 조직, 날짜 등의 개체명의 인식한 예이다.

Jim bought 300 shares of Acme Corp. in 2006.

<ENAMEX TYPE="PERSON">Jim
</ENAMEX> bought
<NUMEX TYPE="QUANTITY">300
</NUMEX> shares of
<ENAMEX TYPE="ORGANIZATION">
Acme Corp.</ENAMEX> in
<TIMEX TYPE="DATE">2006</TIMEX>.

그림 3. 개체명 인식의 예

개체명 인식이 힘든 이유는 새로운 개체명(신조어)이 계속해서 생성되고, 각 분야별로 사용하는 개체명이 다르기 때문에 도메인 의존성이 강하여 오픈 도메인의 경우 성능이 현저히 떨어지기 때문이다.

초기의 개체명 인식 연구는 휴리스틱 및 수동으로 만든 규칙에 의존했지만 방대한 개체명의 양과 계속해서 추가되는 개체명 때문에 2000년대 이후의 규칙 기반 연구는 거의 찾아볼 수 없다. 현재는 지도 학습 기반의 연

구[6]들이 가장 좋은 성능을 보이며 오픈 도메인에 대한 개체명 인식[7]도 계속해서 시도되고 있다. 그리고 대량의 학습 데이터를 생성하는 것은 많은 비용이 드는 작업이므로 적은 양의 학습 데이터를 사용하여 점진적으로 학습 데이터를 확장하는 방식인 반지도 학습 기법을 적용한 연구도 계속되고 있다.

개체명 인식 모듈은 자연어 처리 모듈로부터 출력된 자연어 처리된 데이터를 입력받아 필요한 개체들을 모두 추출한다. 국방 분야에서 쓰이는 개체명에 대해 충분히 알고 있는 전문가가 참여하여 국방 분야에 맞는 개체명의 종류(국가, 조직, 무기명 등)와 개체명을 미리 정의하는 작업과, 일정량 이상의 데이터를 선택하여 그 데이터에 있는 모든 개체명을 사람이 직접 표시하는 학습 데이터 생성 작업이 필요하다. 이렇게 만들어진 학습 데이터를 기반으로 기계 학습 방법을 이용하여 개체명을 인식한다.

개체명 인식 분야에서 지도 학습 기반 방법이 가장 좋은 성능을 보여주고 있으며, 도메인이 한정된 국방 분야의 문서만을 대상으로 개체명을 추출하는 작업이기 때문에 제안하는 방법이 충분한 성능을 보여줄 것이라 기대한다.

3. 관계 추출 모듈(Relation Extraction)

관계 추출은 텍스트 내에서 개체간의 의미적 관계를 탐지하거나 분류하는 작업이다. 관계 추출 모듈은 앞에서 알아본 개체명 인식 과정을 거쳐서 얻어진 개체들 간의 관계를 추출한다. “Bill Gates works at Microsoft Inc.” 문장을 예로 들어 보자. 개체명 인식 과정에서는 ‘Bill′과 ‘Microsoft Inc.’라는 개체명을 추출하고, 관계 추출 단계에서는 두 개체명간의 관계인 Affilation(‘Bill′, ‘Microsoft Inc.′)를 추출해 낸다.

개체명 인식 모듈이 필요한 이유는 단순히 개체명만을 저장해 놓으면 정보로서 활용할 수 없기 때문이다. 위의 예에서 ‘Bill′, ‘Microsoft Inc.’라는 두 개의 개체명만 놓고 보면 특별한 의미가 으며 ‘Affilation′이라는 두 개체간의 관계까지 함께 저장되어 있어야 비로써 완전한 정보로서 가치가 있다.

관계 추출도 최근에는 지도 학습 방법과 반지도 학습 방법을 많이 사용하고 있다. 유전자나 단백질과 질병과의 관계를 나타낸 생물 의학 텍스트[8]나 질문에서 객체들 간의 관계를 파악하기 위한 질의응답 텍스트[9]를 대상으로 연구가 많이 되고 있다.

관계 추출 모듈은 개체명 인식 모듈과 같이 국방 분야에서 필요한 개체명들 간의 관계들을 미리 정의하여 학습 데이터를 만들고 기계학습 기법을 활용하여 관계를 추출한다.

자연어 처리, 개체명 인식, 관계 추출 등의 작업을 묶어서 크게 정보 추출 단계라고 할 수 있는데, 이 단계를 완료하면 구조화 되지 않은 자연어 국방 기사 데이터가 구조화된 데이터로 변환되어 데이터베이스에 저장된다. 정보 추출 단계만 완료되어도 지금까지 수작업으로 각 문서를 찾아서 제작했던 통계 데이터 등을 자동으로 생성할 수 있다.

4. 사용자 요구 분석 모듈(User Needs Analysis)

사용자 분석 모듈은 정보 추출 단계에서 생성된 데이터베이스와 사용자의 요구를 대상으로 수행된다. 사용자가 필요한 정보가 무엇인지 분석하여 정확한 데이터 마이닝을 위한 요구 사항을 생성하고, 그에 필요한 데이터들을 데이터베이스로부터 가져와서 데이터 마이닝 모듈로 보내준다.

5. 데이터 마이닝 모듈(Data Mining)

데이터 마이닝은 대량의 실제 데이터로부터 묵시적이고 전에는 알려지지 않았지만 잠재적으로 유용한 정보를 추출하는 것이다[10]. 데이터 마이닝은 대상이 되는 분야나 텍스트별로 매우 여러 가지 방법이 있지만 크게 4가지로 구분할 수 있다.

1.1 연관성(Association)

동시에 발생하는 사건 그룹 내에서 사건들 사이에 존재하는 친화성 혹은 패턴을 발견하는 작업이다. 예를 들어 IBM의 시장바구니 분석 연구에서는 슈퍼마켓의 소비자들이 구입한 물품 목록을 분석함으로써 콘칩을 구매하는 소비자들의 50%는 콜라도 함께 구매함을 알 수 있었다. 이를 적절히 활용하여 콜라와 콘칩의 결합

상품을 출시하는 마켓팅 등에 활용할 수 있다.

1.2 연속성(Sequence)

순서대로 일어난 데이터를 분석하여 빈도수가 높은 순차 패턴을 찾아내는 작업이다. 각 년도 별 쌀 생산량 데이터와 기후, 경작지 면적, 경작 인원 등의 데이터가 있다면 연속되는 쌀 생산량의 추이를 분석하여 미래의 쌀 생산량을 예측할 수 있을 것이다.

1.3 분류(Classification)

동시에 발생하는 사건 그룹 내에서 사건들 사이에 존재하는 친화성 혹은 패턴을 발견하는 작업이다. 금융기관에서 고객 통계 자료를 바탕으로 대출 대상자를 분류하는 작업이 좋은 예이다. 나이, 연봉, 결혼 유무, 성별 등의 데이터를 바탕으로 새로운 대출 대상자가 입력되었을 때, 대출 가능과 불가능을 분류할 수 있다.

1.4 군집화(Clustering)

유사한 특징을 지닌 몇 개의 소그룹으로 데이터를 분할하는 작업이다. 분류 작업과 유사하나 군집화는 특정 집합이 미리 정의되어 있지 않다는 차이점이 있다. 인터넷 페이지의 Click/View 이벤트를 분석하여 사용자의 특성별로 군집화 하고, 각 사용자 집합과 기사간의 관계를 도출하는 연구[11] 등이 있다.

데이터 마이닝 모듈에서는 사용자 요구 분석 모듈에서 출력한 정보를 바탕으로 연관성, 연속성, 분류, 군집화 등의 방법을 사용하여 사람이 발견하지 못한 새로운 정보를 발견한다. 데이터 마이닝 기법은 매우 다양하기 때문에 이 데이터 마이닝 모듈의 실제 동작 방식은 사용자가 필요로 하는 정보에 크게 달라질 수 있다.

맞춤형 정보 분석 시스템에서 여기까지의 과정은 모두 자동으로 처리된다. 하지만 자동화된 시스템을 거친 이러한 정보가 항상 정확할 수는 없기 때문에 최종적으로 전문가의 확인과 분석이 필요하다. 데이터 마이닝을 통한 새로운 정보와 전문가의 분석이 합쳐져서 시스템을 통한 최종 보고서가 생성된다.

III. 실제 데이터 처리 예제

이번 장에서는 실제 국방 기사 데이터를 대상으로 맞춤형 정보 분석 시스템을 통해 처리한 예를 보여준다.

1. 실제 입력 데이터

[그림 4]와 같은 문장이 실제 입력 데이터로 들어온다. 구조화 되지 않은 자연어 텍스트 데이터이고, 그 내용은 이집트가 록히드마틴사에 전투기와 그 관련 무기의 판매를 요청했다는 내용이다. 이 문장이 자연어 처리 모듈을 거치게 되면 [그림 5]와 같은 형식으로 변환되어 출력된다.

> Egypt has requested the sale of 24 Lockheed Martin F-16C/D Block 50/52 combat aircraft and associated weapons and equipment from the United States, the Defense Security Co-operation Agency (DSCA).

그림 4. 실제 국방 기사 데이터

2. 자연어 처리된 데이터

[그림 5]는 자연어 처리를 거쳐서 품사 부착, 구문 구조 분석이 완료된 데이터이다.

첫 번째 줄을 보면 'Egypt'는 구문 분석 결과 NP(명사구)이며, 품사 부착 결과 NNP(고유명사)라는 사실을 알 수 있다. 마지막의 lem은 이 단어의 어근을 보여준다.

3. 개체명 인식 결과

[그림 6]은 자연어 처리된 결과를 기반으로 문장에 있는 개체명을 추출한 결과이다. 'Egypt'와 'United States'가 위치로 인식되었고, 'Lockheed Martin'은 기관, 'F-16C/D'는 무기명으로 인식된 것을 알 수 있다.

[NP<rel="SBJ" of="s1_1">**Egypt**NNP<lem="Egypt">]
[VP<id="s1_1"> **has**VBZ<lem="have"> **requested**VBN<lem="request">]
[NP<rel="OBJ" of="s1_1"> **the**DT<lem="the"> **sale**NN<lem="sale">]
[PNP
[PP **of**IN<lem="of">]
[NP **24**CD<lem="24"> **Lockheed**NNP<lem="Lockheed"> **Martin**NNP<lem="Martin"> **F-16C**NN<lem="F-16C">]
]
/SYM<lem="&slash;"> [NP<> **D**NNP<lem="D"> **Block**NNP<lem="Block"> **50**CD<lem="50">]
/SYM<lem="&slash;"> [NP<> **52**CD<lem="52"> **combat**NN<lem="combat"> **aircraft**NN<lem="aircraft">]
andCC<lem="and"> [VP<id="s1_2"> **associated**VBN<lem="associate">]
[NP<rel="OBJ" of="s1_2"> **weapons**NNS<lem="weapon"> **and**CC<lem="and"> **equipment**NN<lem="equipment">]
[PNP
[PP **from**IN<lem="from">]
[NP<> **the**DT<lem="the"> **United**NNP<lem="United"> **States**NNP<lem="States">]
]
,comma<lem=","> [NP<> theDT<lem="the"> **Defense**NNP<lem="Defense"> **Security**NNP<lem="Security">
Co-operationNNP<lem="Co-operation"> **Agency**NNP<lem="Agency">]
(openparen<lem="("> [NP<> **DSCA**NNP<lem="DSCA">]
)closeparen<lem=")">

그림 5. 자연어 처리된 국방 기사 데이터

[LOC/Egypt] has requested the sale of 24 **[ORG/Lockheed Martin]** **[WEAP/F-16C/D]** Block 50/52 combat aircraft and associated weapons and equipment from the **[LOC/United States]** , the **[ORG/Defense Security Co-operation Agency]** (**[ORG/DSCA]**).

그림 6. 개체명 인식 결과

4. 관계 추출과 데이터베이스화

[표 1]은 인식된 개체명으로부터 그 관계를 추출하여 데이터베이스화 결과를 나타낸다. 이 표는 객체들의 관계만을 나타낸 테이블이며, 그 외 객체의 종류, 부가적인 정보가 있는 테이블들도 함께 저장될 것이다.

표 1. 데이터베이스화 된 개체명과 관계

Relation	Entity1	Entity2
Sales Request	Egypt	Lockheed Martin
Product	Lockheed Martin	F-16C/D
Nationality	Lockheed Martin	United States

5. 데이터 마이닝

정보 추출 단계에서 생성된 데이터베이스를 대상으로 데이터 마이닝을 수행하면 다음과 같은 질문들에 답할 수 있다. “Egypt의 우호국들은?”, “Egypt에서 추후 구매할 예상 무기는?” 무기 판매 요청, 국가 원수의 방문, 원조 등의 여러 정보들을 종합하여 ‘Egypt’의 우호국에 대한 정보를 찾아낼 수 있고, 지금까지 ‘Egypt’가 구매한 연도별 무기 목록, 무기들의 연관성을 고려하면 추후 구매할 예상 무기들의 목록을 얻을 수 있다.

IV. 결론

지금까지 맞춤형 정보 분석을 위해 필요한 기술들과 국방 기사 데이터를 대상으로 맞춤형 정보 분석을 자동으로 수행하는 시스템에 대하여 알아보았다 . 많은 전문가가 투입되어 계속해서 생성되는 모든 데이터를 직접 확인하여 정보 분석을 수행하는 것이 가장 좋은 방법이지만, 제한된 인력과 많은 비용 때문에 현실적으로 불가능한 방법이다.

본 논문에서 제안하는 맞춤형 정보 분석 시스템은 현재 연구되어 있는 기술들을 이용하여 정보 분석의 일정 부분을 자동으로 수행함으로서 정보 분석에 소요되는 시간과 비용을 줄이고 그 신뢰도를 높일 수 있는 시스템이다. 실제로 이 시스템을 구현하여 사용한다며 국방 정보 분석에 많은 도움이 될 것이라 기대한다.

참 고 문 헌

[1] J. Gimenez and L. Marquez, "Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited", In Proceedings of Recent Advances in Natural Language Processing, pp.153-163, 2003.

[2] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging", In Proceedings of Association for Computational Linguistics, pp.744-751, 2007.

[3] M. Collins, "Head-driven statistical models for natural language parsing", Journal of Association for Computational Linguistics, Vol.29, No.4, pp.589-638, 2003.

[4] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov, and E. Marsi, "MaltParser: A language-independent system for data-driven dependency parsing", Journal of Natural Language Engineering, Vol.13, No.2, pp.95-136, 2007.

[5] M. Marneffe, B. Maccartney and C. Manning, "Generating Typed Dependency Parses from Phrase Structure Parses", In Proceedings of International Conference on Language Resources and Evaluation, pp.449-454, 2006.

[6] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons", In Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics, pp.188-191, 2003.

[7] R. Evans, "A framework for named entity recognition in the open domain", In Proceedings of Recent Advances in Natural Language Processing, pp.267-276, 2003.

[8] H. Chun, Y. Tsuruoka, J. Kin, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning", In Pacific Symposium on Biocomputing, pp.4-15, 2006.

[9] E. Hovy, U. Hermjakob and D. a. Ravichandran, "A question/answer typology with surface text patterns", In Proceedings of Human Language Technology, pp.247-251, 2002.

[10] W. J. Frawley, G. Piatetsky-Shapir, and C. J. Matheus, "Knowledge Discovery in Databases: An Overview", AI Magazine, Vol.13, No.3, pp.57-70, 2003.

[11] W. Chu, S. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah, "A case study of behavior-driven conjoint analysis on Yahoo!: front page today module", In Proceedings of Knowledge Discovery and Data Mining., pp.1097-1104, 2009.

[12] C. D. Manning and H. Schuetze, Foundations of Statistical Natural Language Processing, The

MIT Press, 1999.

저 자 소 개

최 중 환(Joong-Whoan Choi) 정회원

- 1987년 2월 : 부산대학교 금속공학과 졸업(학사)
- 1992년 8월 : 부산대학교 금속공학과 졸업(석사)
- 2000년 2월 : 부산대학교 금속공학과 졸업(박사)
- 1988년 3월 ~ 현재 : 국방기술품질원

<관심분야> : 금속 열처리, 표면처리, 정보기술, 정보분석

임 채 오(Chea-O Lim) 정회원

- 1995년 2월 : 국방대학교 무기체계공학과 졸업(석사)
- 1989년 9월 ~ 현재 : 국방기술품질원

<관심분야> : STEP, CAD/CAM, PLM