대화형 개인 비서 시스템을 위한 하이브리드 방식의 개체명 및 문장목적 동시 인식기술

이창수○, 고영중
동아대학교
blue772001@gmail.com, youngjoong.ko@gmail.com

A Simultaneous Recognition Technology of Named Entities and Objects for a Dialogue Based Private Secretary Software

ChangSu Lee○, YoungJoong Ko
Donga University, Computer Engineering

요 약

기존 대화시스템과 달리 대화형 개인 비서 시스템은 사용자에게 정보를 제공하기 위해 앱(APP)을 구동하는 방법을 사용한다. 사용자가 앱을 통해 정보를 얻고자 할 때, 사용자가 필요로 하는 정보를 제공해주기 위해서는 사용자의 목적을 정확하게 인식하는 작업이 필요하다. 그 작업 중 중요한 두 요소는 개체명 인식과 문장목적 인식이다. 문장목적 인식이란, 사용자의 문장을 분석해 하나의 앱에 존재하는 여러 정보 중 사용자가 원하는 정보(문장의 목적)가 무엇인지 찾아주는 인식작업이다. 이러한 인식시스템을 구축하는 방법 중 대표적인 방법은 사전규칙방법과 기계학습방법이다. 사전규칙은 사전정보와 규칙을 적용하는 방법으로, 시간이 지남에 따라 새로운 규칙을 추가해야하는 문제가 있으며, 규칙이 일반화되지 않을 경우 오류가 증가하는 문제가 있다. 또 두 인식작업을 파이프라인 방식으로 적용 할 경우, 개체명 인식단계에서의 오류를 가지고 문장목적 인식단계로 넘어가기 때문에 두 단계에 걸친 성능저하와 속도저하를 초래할 수 있다. 이러한 문제점을 해결하기 위해 우리는 통계기반의 기계학습방법인 Conditional Random Fields(CRF)를 사용한다. 또한 사전정보를 CRF와 결합함으로써, 단독으로 수행하는 CRF방식의 성능을 개선시킨다. 개체명과 문장목적인식의 구조를 분석한 결과, 비슷한 자질을 사용할 수 있다고 판단하여, 두 작업을 동시에 수행하는 방법을 제안한다. 실험결과, 사전규칙방법보다 제안한 방법이 문장단위 2.67% 성능개선을 보였다.

주제어: 개체명 인식, 하이브리드, Conditional Random Fields, 문장목적인식, 대화시스템

1. 서론

모바일기술의 발달로, 대화시스템이 기존의 오프라인 대화시스템으로부터, 실시간 정보획득 및 질의응답을 목적으로 한 온라인 대화시스템으로 변화되고 있다. 온라인 대화시스템의 하나인 대화형 개인 비서 시스템의 가장 큰 장점은 앱(APP)을 구동할 수 있는데 있다. 기존의 대화시스템은 규칙이나 학습을 통해 정해진 질문에 정해진 대답만을 할 수 있는데 반해, 이 시스템은 앱을 통해 사용자의 질문에 따라 사용자가 원하는 화면을 보여주며, 실시간으로 정보를 제공한다.

사용자가 앱을 통해 정보를 얻고자 할 때, 시스템에서 사용자가 필요로 하는 정보를 제공해주기 위해서는 사용자의 목적을 정확히 인식하는 작업이 중요하다. 이러한 작업 중 중요한 두 가지 요소는 개체명 인식과 문장목적 인식이다.

개체명 인식은 질의응답시스템과 정보검색 분야에서 유용하게 사용되고 있는 정보추출의 한 단계이다. 개체명은 인명, 지명, 조직명, 시간, 날짜, 화폐 등의 고유명사이며, 개체명인식은 문장에서 개체명을 식별하고 식별된 개체명의 종류를 결정하는 작업이다. 이 작업은 문장에서 중요한 핵심어를 추출해 문장의 의미를 파악하는데 도움을 준다.

개체명 인식 방법은 크게 세 가지 방법으로 나눌 수 있다. 첫 번째 방법은 규칙기반 개체명 인식[1][2][3]이며, 규칙기반 개체명 인식은 사전정보나 규칙만을 사용하기 때문에 다음과 같은 문제점을 가지고 있다.

1. 사전정보나 규칙이 포함되어 있지 않은 문장 인식 문제.
2. 시간이 지남에 따라, 규칙을 지속적으로 추가해줘야 하는 문제.
3. 규칙이 일반화되지 않을 경우 많은 오류를 유발 하는 문제.

두 번째 방법은 통계기반의 기계학습 방법이다[4][5][6][7]. 통계기반의 방법을 사용하기 위해선 양질의 말뭉치가 필요하며, 자질을 색출하는 작업을 거쳐야 한다. 마지막으로 규칙기반과 기계학습을 결합한 방법도 있다[8][9].

문장목적인식이란, 앱 구동을 통해 실행된 하나의 앱

본 연구는 산업자원통상부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]

에서는 여러 정보가 있을 수 있는데, 그 정보들 중 사용자가 필요로 하는 정보를 제공하기 위해 문장의 목적을 찾는 작업이다. 예를 들어, 날씨 앱을 구동 했을 때, 사용자는 오늘 비가 올 것인지, 기온이 어떻게 되는지, 날씨가 맑은지, 바람이 많이 부는지 등 여러 정보 중 원하는 하나의 정보가 있을 것이다. 문장목적 인식은 사용자의 문장을 분석해, 위와 같은 여러 정보 중 사용자가 필요로 하는 정보가 무엇인지 찾아 주는 인식작업이다. 기존의 대화시스템에서는 이러한 문장목적을 인식하는 시스템이 존재하지 않았다. 하지만 앱을 구동하는 대화형 개인 비서 시스템에서는 문장목적을 인식하는 것은 사용자가 원하는 정보를 앱에서 찾아 사용자에게 정확히 제공해주기 위해 반드시 필요한 인식작업이다.

대화형 개인 비서 시스템에선 개체명 인식, 문장목적 인식을 통해 앱에서 사용자에게 제공해야 할 정보가 무엇인지 파악한다. 예를 들어, 사용자가 "오늘 기온"에 대한 정보를 얻고자 할 때, 사용자는 "오늘 기온이 어때?" 라는 질의를 시스템에게 보내고, 개체명 인식 작업을 통해 "오늘"이라는 개체명을 인식한 후, 문장목적 인식 작업을 통해 문장의 목적이 "기온"에 대한 정보라는 것을 파악한다. 이러한 과정을 거쳐 시스템은 날씨앱을 실행시켜, 오늘의 기온에 대한 정보를 앱에서 찾아 사용자에게 제공해주는 것이다.

본 논문에서는 개체명 및 문장목적 인식을 위해 통계기반의 기계학습기법인 CRF와 사전정보를 함께 사용하는 하이브리드 방식을 제안한다. 또한 개체명과 문장목적을 인식하는 방법의 구조를 분석한 결과, 서로 비슷한 자질로 분류가 가능하다고 판단하여 서로 다른 두 개의 인식 시스템을 단 한 번의 기계학습을 통해 동시에 인식하는 방법을 제안한다. 2장에서는 관련연구에 대해 살펴보고, 3장에서는 개체명 인식 시스템, 문장목적 인식 시스템, 개체명 및 문장목적 동시 인식 시스템에 대해 각각 사전규칙, CRF, 하이브리드방식으로 나누어 설명하고, 4장에서는 3장에서 제안한 하이브리드 시스템을 사전규칙 시스템과 CRF만을 사용한 시스템과 비교하여 하이브리드 시스템의 유용성을 살펴보며, 본 논문에서 제안한 하이브리드 기반의 개체명 및 문장목적 동시 인식기술이 사전규칙방식이나 단독CRF방식을 사용해 두 인식을 수행한 것보다 더 합리적인 것을 살펴본다. 5장에서는 결론 및 향후 연구과제에 대하여 기술한다.

2. 관련 연구

개체명 인식은 질의응답시스템과 정보검색 분야에서 유용하게 사용되고 있는 정보 추출의 한 분야로서 문서나 문장 내에서 개체명을 추출하고 추출된 개체명의 종류를 식별 하는 작업을 말한다. 개체명 인식에 관한 연구는 1990년대에 정보추출(Information Extraction)의 목적으로 개최되었던 Message Understanding Conference (MUC)에서 본격적으로 연구되기 시작해, MUC 이후 개체명에 대한 연구가 꾸준히 진행 되어왔으며, Conference on Computational Natural Language Learning 2002(CoNLL 2002)와 CoNLL2003을 통해서 더욱 많은 발전이 있었다[10]. 개체명 인식은 크게 3가지 방법으로 연구되었다. 첫 번째는 규칙 기반 방법이며 이 방법에서는 주로 정규표현식[3]이나 자연어 특징을 이용한 규칙과 사전정보[2]를 사용했다. 두 번째로 통계기반의 기계학습 방법이며, 대표적인 방법으로 Hidden Markov Model, Maximun Entropy Model, Conditional Random Fields, Decision Tree 등이 있다[4][5][6][7]. 마지막으로 규칙기반과 기계학습을 함께 사용한 하이브리드 방법도 연구되었다[8][9]. 개체명 인식 연구가 시작된 1990년대에는 대부분 영어만을 대상으로 개체명 인식 연구가 이루어졌지만 최근에서 영어뿐만 아니라 한국어[5], 일본어, 중국어 등 다양한 언어에 대해서 개체명 인식 시스템이 연구되고 있다.

3. 개체명 및 문장목적 인식

우리는 사전정보와 CRF를 결합하는 하이브리드 방법을 통해 개체명 및 문장목적 인식의 성능을 높이는 방법을 시도하며, 개체명과 문장목적 인식이라는 각각의 작업을 동시에 수행하는 방법을 제안한다.

대화형 개인 비서 시스템에서 사용자의 목적을 파악해 원하는 정보를 실시간으로 제공 하기위해서는 개체명과 문장목적을 인식하는 작업이 필요하다. 이에 따라 본 논문은 다음의 3가지 인식작업을 나눠서 살펴 본 후, 개체명과 문장목적을 동시에 인식하는 방법이 합리적인 접근 방법임을 보인다.

1.개체명 인식
2.문장목적 인식
3.개체명과 문장목적의 동시 인식

대화형 개인 비서 시스템은 6개의 도메인을 가지며, 이 6개의 도메인에서 출현하는 개체명과 문장목적을 인식 대상으로 한다.

표1. 도메인 종류

교통	날씨	시계	알람	일정	환율

3.1 개체명 인식

개체명의 종류는 표2와 같이 8가지로 분류 된다.

표2. 개체명 종류

인명 (Person)	지명 (Location)	날짜 (Date)	시간 (Time)
반복 (Cycle)	타이틀 (Title)	통화 (Currency)	숫자 (Number)

개체명 종류에서 반복은 "매년, 매주, 매월" 같은 단어이며, "논문 계획, 미팅 일정, 점심 약속"같은 단어를 타이틀이라고 명명한다. 개체명 인식 시스템은 비교를 위해 사전규칙, CRF, 하이브리드 방식으로 각각 시스템을 구축한다. 먼저, 사전규칙기반은 각 도메인별로 사전과 규칙을 다르게 적용하여, 도메인별로 사전이나 규칙이 중복되어 오류를 유발할 수 있는 부분을 최대한 줄였다. 다음은 사전 정보와 규칙적용 방식을 보여준다.

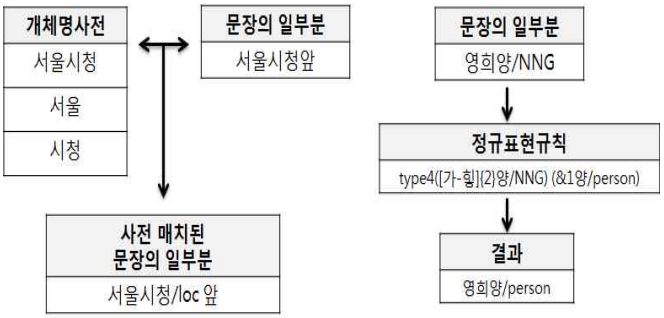

그림1. 사전매치방법(왼쪽)과 규칙적용방법(오른쪽)

규칙은 여러 가지 타입에 유연하게 적용할 수 있도록, 어휘정보 및 형태소정보, 정규표현 등 5가지의 타입규칙을 사용하였다. 또한 사전정보는 최장길이일치법을 적용했다. 하이브리드 방식에서는 개체명 인식 시스템의 성능 향상을 위해 CRF에 사전정보를 결합함으로써 개체명 인식 시스템의 성능을 높이는 방법을 시도했다. 또한 사전정보와 CRF를 각각 단독으로 사용하는 시스템을 구축해 하이브리드 방식이 합리적인지 평가했다. 그림2는 하이브리드 기반의 개체명 인식시스템의 구조도이다.

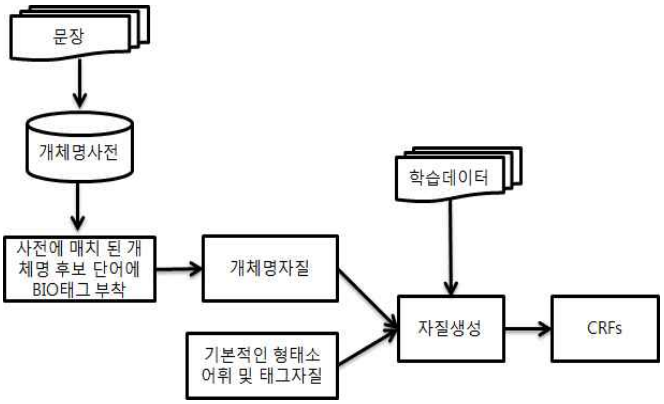

그림2. 하이브리드기반의 개체명 인식시스템 구조도

사전정보를 이용해 사전에 매치된 단어는 개체명 후보로 인식하며, 개체명인식 중 날짜와 시간은 정해진 템플릿이 존재하기 때문에 기존의 규칙시스템에서 사용한 방법을 그대로 사용한다. 그리고 사전에 매치된 단어는 그림3과 같이 BIO태그를 부착해 CRF를 사용할 때 개체명사전자질을 사용할 수 있도록 구성했다.

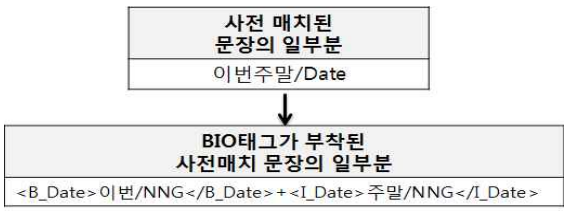

그림3. BIO태그가 부착된 사전정보

3.2 문장목적 인식

문장목적의 종류는 6개의 도메인에서 28개의 세부목적이 존재한다. 대표적인 문장목적 종류는 알람(Alarm), 지역시간(LocationTime), 버스스케줄(busSchedule), 날씨정보(weatherInfo)등이 있다.

문장목적 인식에서는 문장의 목적을 파악하기 위해 다음과 같은 두 가지 학습방법을 사용할 수 있다.

1.문장 전체를 학습시켜 목적을 찾아내는 방법
2.문장에서 중요한 실마리가 되는 부분의 구간을 정해 그 구간이 존재하는지 확인해 목적을 찾아내는 방법.

문장전체를 학습시키는 방법은 개체명과 문장목적 인식을 수행하기 위해 서로 다른 기계학습 기법을 사용해야 하는 단점이 있다. 그 이유는, 인식 단위의 차이 때문이다. 즉, 개체명 인식에서는 형태소별, 목적인식에서는 문장별로 분류하게 된다. 실시간 시스템에서는 속도가 매우 중요하기 때문에, 우리는 두 가지 인식을 동시에 수행하기 위한 기반을 마련하기 위해 실마리구간 탐색방법을 사용했다.

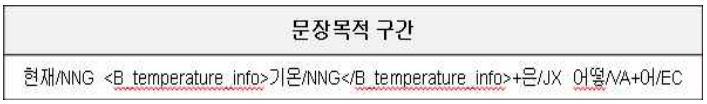

그림4. 문장의 실마리 구간

그림4에서는 문장의 목적이 기온정보인 경우 문장에서 기온정보의 실마리 구간을 "기온"라는 단어를 실마리라고 정해 이 문장이 기온정보에 대한 목적을 가진 문장인지를 식별하게 해준다.

문장목적 인식 사전규칙시스템은 개체명 인식과 같은 방법으로 각 도메인별 사전과 규칙을 적용시켜 오류를 최소화 시켰다. 통계기반의 문장목적 인식시스템도 마찬가지로 사전정보와 CRF를 결합하는 하이브리드 방식을 통해 성능을 높이는 방법을 시도했다. 또한 사전정보와 CRF를 각각 단독으로 사용하는 시스템을 구축해 하이브리드 방식이 합리적인지 평가했다. 그림5는 문장목적인식의 구조도이다.

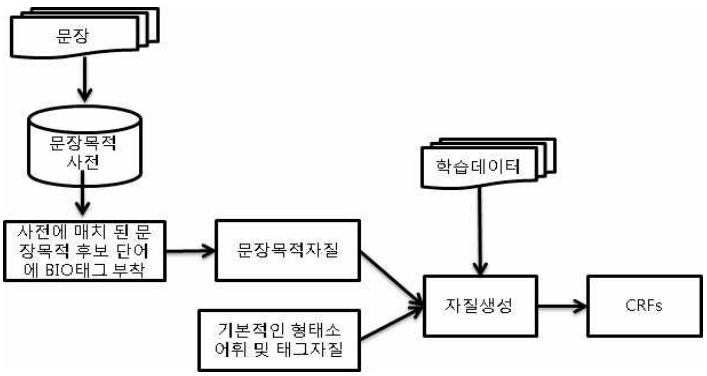

그림5. 하이브리드기반의 문장목적 인식시스템 구조도

3.1절의 개체명 인식과 같은 단계를 거쳐 목적구간을 찾아 문장의 목적 인식작업을 수행하도록 구축했다.

3.3 개체명 문장목적 동시 인식

3.1과 3.2절에서 개체명과 문장목적을 인식하는 구조

도를 보였다. 두 시스템의 비교결과, 문장목적인식에 구간정보를 사용함으로써, 개체명 인식과 같은 시스템구조를 사용할 수 있으며, 따라서 우리는 두 인식 시스템이 서로 비슷한 자질로 분류가 가능할 것이라고 판단했다. 이에 따라 개체명과 문장목적을 동시에 인식하는 시스템을 구현 하는 것은 합리적인 방법일 것이라고 생각했다. 우리가 생각한 방법이 올바른지 판단하기 위해 사전규칙, CRF, 하이브리드기반 시스템을 각각 구현해 비교했다.

사전규칙기반 시스템은 개체명과 문장목적인식을 파이프라인 방식으로 수행한다. 그림6은 사전규칙기반 시스템의 구조도이다.

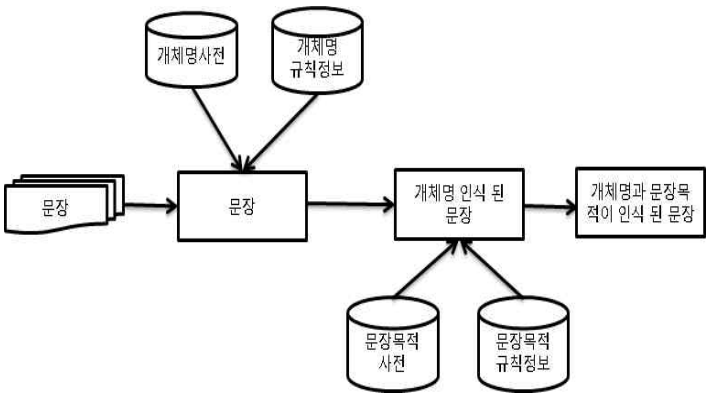

그림6. 개체명과 문장목적 인식을 파이프라인 방식으로 수행하는 사전규칙기반 시스템의 구조도

사전규칙기반은 파이프라인 방식을 수행함에 있어, 다음과 같은 문제점을 가지게 된다.

1.개체명 단계에서 생긴 오류를 가지고 문장목적 단계로 넘어가기 때문에 두 단계에 걸친 성능저하.

2.두 단계의 순차적 수행으로 인한 속도저하.

반면, 본 논문에서 제안한 시스템은 두 인식 시스템이 같은 구조를 갖는다는 것을 바탕으로 개체명과 문장목적을 동시에 수행함으로써 위의 문제점을 해결한다. 동시인식 시스템 또한 사전정보와 CRF를 결합하는 하이브리드 방식을 통해 성능을 높이는 방법을 시도했다. 또 CRF를 단독으로 사용하는 시스템을 구축해 제안한 방법이 합리적인지 평가했다. 그림7은 제안한 하이브리드기반 시스템의 구조도이다.

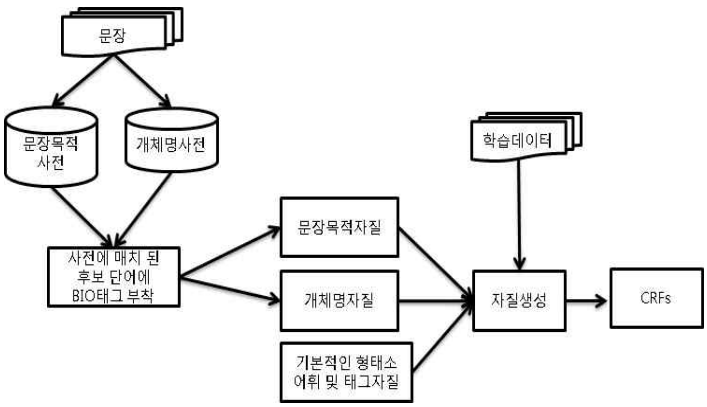

그림7. 개체명과 문장목적을 동시에 인식하는 하이브리드기반 시스템의 구조도

하이브리드기반 시스템은 사전정보를 활용해, 사전에 매치된 개체명후보와 문장목적 구간후보를 인식해, CRF를 사용할 때 사전자질을 사용할 수 있도록 구성했다.

3.4 자질집합

자질 집합은 형태소어휘, 형태소태그, 어절 내 자질등 기본적으로 사용하는 자질집합과 함께, 인식 성능을 높이기 위해 본 논문에서 구축한 사전자질, 개체명자질, 문장목적 자질을 추가하였다. 표3는 전체 자질집합을 보여준다.

표3. 시스템에 사용된 자질집합

형태소어휘태그 자질	1.형태소어휘/태그
형태소어휘자질	2.w_0/w_1 3.w_{-1}/w_0 4.$w_{-1}/w_0/w_1$, 5.$w_0/w_1/w_2$ 6.$w_{-2}/w_{-1}/w_0$ 7.$w_{-1}/w_0/w_1/w_2$ 8.$w_{-2}/w_{-1}/w_0/w_1$ 9.$w_{-2}/w_{-1}/w_0/w_1/w_2$
형태소태그자질	10.p_0/p_1 11.p_{-1}/p_0 12.$p_{-1}/p_0/p_1$ 13.$p_0/p_1/p_2$ 14.$p_{-2}/p_{-1}/p_0$ 15.$p_{-1}/p_0/p_1/p_2$ 16.$p_{-2}/p_{-1}/p_0/p_1$ 17.$p_{-2}/p_{-1}/p_0/p_1/p_2$
어절 내 자질	18.형태소 어절 내 위치 19.형태소태그/어절길이
사전 자질	20.형태소어휘/태그 + 레이블정보
개체명자질	21.현재 위치의 앞에 나온 모든 개체명 정보 22.현재 위치의 앞에 나온 모든 개체명의 시퀀스정보(NULL 포함) 23.형태소어휘/태그와 문장에 존재하는 모든 개체명정보 24.형태소어휘/태그와 문장에 존재하는 모든 개체명의 시퀀스정보 (NULL포함)
문장목적자질	25.현재 위치의 앞에 나온 모든 문장목적 정보 26.현재 위치의 앞에 나온 모든 문장목적의 시퀀스정보(NULL 포함) 27.형태소어휘/태그와 문장에 존재하는 모든 문장목적정보 28.형태소어휘/태그와 문장에 존재하는 모든 문장목적의 시퀀스정보 (NULL포함)

시스템에 사용된 자질집합은 총 28개로 구성된다. 형태소어휘자질은 현재형태소어휘(w_0)를 중심으로 이전/이후의 연속된 형태소어휘이다. 형태소태그자질도 마찬가지로 현재형태소태그(p_0)를 중심으로 이전/이후의 연속된 형태소태그이다. 형태소의 어절 내 위치자질은 형태소가 어절의 시작, 중간, 끝을 나타내는 정보이며, 형태소태그/어절길이자질은 형태소태그와 형태소를 포함하는 어절의 길이정보이다. 마지막으로 사전자질은 형태소어휘/태그와 형태소의 레이블정보이다. 레이블은 사전매치를 통해 후보가 된 개체명이나 문장목적에 BIO태그가 부착된 정보이다.

4. 실험결과

모든 실험결과는 정확도(Accuracy)를 측정해 성능을 평가하며, 개체명과 문장목적을 동시에 인식하는 시스템은 문장 내에 모든 개체명과 함께 문장의 목적을 정확히 인식했을 경우에만 맞는 문장이라고 간주한다. CRF 및 하이브리드기반 시스템에 사용한 자질 중 실험결과 가장 좋은 성능을 보이는 자질집합을 각 실험결과의 하위에 기재했다. 또한 모든 자질은 표3에 기재한 자질만을 사용했으며, 각 실험에서 사용한 자질을 쉽게 보여주기 위해 자질 별로 번호를 부여해, 자질번호 나열을 통해 사용한 자질을 보여주는 방식으로 구성했다.

4.1 실험데이터

개체명과 문장목적 인식을 위해 ETRI개체명사전을 사용했으며, 학습데이터를 통해 문장목적구간사전을 구축했다. 또한 말뭉치는 전문적으로 태깅을 하는 전문가를 통해 6개의 도메인에서 2972개의 문장을 태깅한 것을 사용했다. 표4는 6개의 도메인에 대한 문장의 분포이다.

표4. 도메인에 따른 말뭉치 분포

교통	날씨	시계	알람	일정	환율
355	603	246	658	895	215

본 논문에서는 학습데이터에 1925개의 문장을 사용했고, 테스트데이터는 1047개의 문장을 사용했다.

4.2 개체명 인식 결과

개체명 인식은 4개의 버전으로 성능을 비교했다.
표5는 개체명 인식에 대한 실험결과이다.

표5. 개체명 인식 성능

개체명인식 방법	문장 단위 정확도(Accuracy)
사전	86.72%
사전+규칙	93.50%
CRF	91.88%
하이브리드 (사전+CRF)	93.50%

표5에서 개체명 인식 성능은 사전규칙기반 시스템과 하이브리드기반 시스템이 높은 것을 확인할 수 있다. 하이브리드기반 시스템에서는 CRF에 개체명자질과 사전자질을 추가함으로써, CRF를 단독으로 사용한 방법보다 높은 성능을 낼 수 있음을 보였다. 서로 비슷한 의미를 지닌 문장은 같은 개체명 종류가 나올 것이라는 가정을 통해 구축한 개체명사전자질의 타당성을 실험을 통해 입증했으며, 하이브리드 방식에서 개체명인식 성능을 개선하는데 기여했음을 알 수 있다.

개체명 인식 방법 중 CRF자질은 형태소어휘/태그자질(1), 형태소어휘자질(2,3,4,9), 형태소태그자질(10,11,12,13,14,17), 어절 내 자질(18,19)을 사용했다.

하이브리드 방법은 형태소 어휘/태그자질(1), 형태소어휘자질(2,3,4,5,6,7,8,9), 형태소태그자질(10,11,12,13,14,15,16,17), 어절 내 자질(18,19), 사전자질(20), 개체명자질(21,22,23,24)을 사용해 성능을 평가했다.

4.3 문장목적 인식 결과

문장목적 인식 또한 4개의 버전으로 성능을 비교했다. 표6은 문장목적인식에 대한 결과이다.

표6. 문장목적 인식 성능

문장목적 인식방법	문장 단위 정확도(Accuracy)
사전	80.80%
사전+규칙	95.41%
CRF	99.71%
하이브리드 (사전+CRF)	99.31%

문장목적 인식은 사전규칙기반 시스템보다 CRF와 하이브리드기반 시스템을 수행한 결과가 더 높은 성능을 보이는 것을 결과를 통해 확인 할 수 있었다. 그리고 문장목적 인식은 문장목적사전자질을 추가했을 때보다 기본적인 자질만을 사용했을 때, 더 높은 성능을 보이는 것을 확인했다. 그 이유는 문장목적인식은 구간을 통해 문장의 목적을 식별하기 때문에, 완전히 같은 단어와 단어집합, 태그정보가 반복적으로 구간에 나타나는 경향이 많았다. 그런 이유로, 기본적인 자질을 사용했을 때 성능이 더 높은 것으로 분석되었다.

문장목적 인식 방법 중 CRF자질은 형태소어휘/태그자질(1), 형태소어휘자질(2,3,4,5,6,7,8,9), 형태소태그자질(10,11,12,13,14,15,16,17), 어절 내 자질(18,19)을 사용했다.

하이브리드 방법에서는 사전자질 및 문장목적자질을 추가 할 때 마다 인식성능이 떨어지는 것이 확인되었다. 그러므로 CRF에서 사용한 자질 중 사전자질(20)만을 추가한 결과를 사용해 성능을 평가했다.

4.4 개체명 문장목적 동시 인식 결과

개체명 문장목적 동시 인식방법은 3가지 버전으로 성능을 비교했다. 표7은 개체명 문장목적 동시인식에 대한 결과이다.

표7. 개체명 및 문장목적 동시 인식 성능

개체명 및 문장목적 인식방법	문장 단위 정확도(Accuracy)
사전+규칙	88.92%
CRF	91.50% (개체명 91.69%, 문장목적 98.80%)
하이브리드 (사전+CRF)	91.59% (개체명 93.12%, 문장목적 98.09%)

사전규칙기반 시스템은 파이프라인을 통해 수행되기

때문에 개체명과 문장목적 인식성능을 따로 기재하지 않고, 전체 성능만을 기재했다. 사전규칙기반 시스템은 오류가 축적됨에 따라 개체명 인식, 문장목적 인식을 각각 수행할 때 보다 문장 단위 정확도가 많이 떨어진 것을 확인할 수 있었다.

하이브리드기반 시스템에서는 기본적인자질과 사전자질을 사용해, 동시에 두 개의 작업을 수행함으로써, 사전규칙기반 시스템보다 2.67%의 성능이 향상된 것을 확인할 수 있었다. 개체명 문장목적 동시 인식 실험에서 확인할 수 있었던 것은 문장목적은 기본적인자질(시퀀스자질)이 높은 성능을 보이는데 반해, 개체명은 기본적인자질과 함께 개체명사전자질을 사용했을 때 더 높은 성능을 보였다. CRF를 단독으로 사용한 시스템에서는 개체명인식 성능은 상대적으로 낮은 성능을 보였지만, 문장목적 인식결과가 상당히 높은 성능을 보여, 문장 단위 성능이 높게 나온 것을 확인했다.

동시 인식 방법 중 CRF자질은 형태소어휘/태그자질(1), 형태소어휘자질(2,3,4,5,6,7,8,9), 형태소태그자질(10,11,12,13,14,15,16,17), 어절 내 자질(18,19)을 사용했다.

하이브리드 방법은 형태소어휘/태그자질(1), 형태소어휘자질(2,3,4,5,6,7,8,9),형태소태그자질(10,11,12,13,14,15,16,17), 어절 내 자질(18,19), 사전자질(20), 개체명자질(21,22,23,24)을 사용해 성능을 평가했다.

5. 결론

본 논문에서는 대화형 개인 비서 시스템에서 중요한 작업인 개체명 인식과 문장목적 인식을 동시에 수행하는 방법을 제안했다. 6개의 도메인에서 8개의 개체명과 28개의 문장목적을 대상으로 개체명과 문장목적 분류를 수행한 결과, 사전규칙기반의 파이프라인방식을 통해 두 인식을 수행한 성능보다 CRF와 사전자질을 이용해 하이브리드방식으로 두 인식을 동시에 수행한 결과가 문장단위 2.67%의 성능이 향상된 것을 확인 할 수 있었다.

향후 연구로는 문장목적구간이 학습데이터에서는 문장당 한 구간만 존재했지만, 서로 다른 목적을 가진 구간이 한 문장에 2개 이상 출현했을 경우 여러 목적구간 중 가장 확률이 높은 것을 선택하는 방법을 연구할 것이며, 대화형 개인 비서 시스템에서는 개체명인식과 문장목적 인식 이외에 형태소분석, 화행분석, 도메인인식 등 여러 절차가 존재하는데, 그 절차 중 도메인인식은 본 논문에서 제안한 방법과 결합이 가능 할 것으로 생각된다. 그러므로 개체명, 문장목적, 도메인 인식을 동시에 수행하는 방법을 연구할 것이다.

참고문헌

[1] Krupka, G.R. and Hausman, K. "Description of the netowl text extraction system as using for MUC-7." In proceedings of the Seventh Message Understanding Conference(MUC-7) 1998.

[2] 이경희, 이주호, 최명석, 김길창, “한국어 문서에서 개체명 인식에 관한 연구” 한국정보과학회 언어공학연구회 학술발표 논문집16 pp.40-45, 2004

[3] Mesfar, S. "Named Entity Recognition for Arabic Using Syntactic Grammars." 12th International Conference on Application of Natural Language to Information Systems, pp. 305-316, 2007.

[4] Nadeau, D. and Sekine, S. "A Survey of Named Entity Recognition and Classfication" Lingvisicae Investigationes, 30(1), pp.3-26, 2007.

[5] 이창기, 황인규, 오효정, 임수종, 허정, 이충희, 김현진, 왕지현, 장명길 “Conditional Random Fields를 이용한 세부 분류 개체명 인식” 한국정보과학회 언어공학연구회 학술발표 논문집, pp.268-272, 2006.

[6] Lafferty, J. McCallum, A.Pereira, F., "Conditional random fields : Probabilistic models for segmenting and labeling sequence data", ICML, pp.282-2289, 2001.

[7] Ratnaparkhi, A., "A Simple Introduction to Maximum Entropy Models for Natural Language Processing." University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-97-08, 1997.

[8] Petasis, G., Vichot, F., Wolinskim, F., Paliouras, G., Karkaletsis, V. and Spyropoulos, C. D. "Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems." Proceeding Conference of Association for Computational Linguistics, pp.426-433, 2001.

[9] Mai Mohamed Oudah, Khaled Shaalan "A Pipeline Aribic Named Entity Recognition Using a Hybrid Approach" COLING, pp.2159-2176, 2012.

[10] Kim Sang, E. F. T., de meulder, F., "Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition" CoNLL 2003.