대화형 개인 비서 시스템의 언어 인식 모듈을 위한 개체명 및 문장목적 동시 인식 방법

(A Simultaneous Recognition Method of Named Entities and Objects for the Language Understanding Module of a Dialogue-based Private Secretary System)

이 창 수 † 고 영 중 ††
(Changsu Lee) (Youngjoong Ko)

요 약 대화형 개인 비서 시스템은 기존의 대화 시스템과 달리 앱(App)을 이용하여 사용자에게 정보를 실시간으로 제공하는 시스템이다. 대화형 개인 비서 시스템에서의 SLU(Spoken Language Understanding) 작업은 도메인, 개체명, 문장목적, 동작, 화행 인식으로 나누어진다. 본 논문은 대화형 개인 비서 시스템에서의 SLU 작업들 중 개체명과 문장목적을 동시에 인식하는 방법을 연구한다. 기존 시스템은 사전-규칙 기반 방법 이용하여 인식 작업을 수행하는데, 이 방법은 몇 가지 문제점이 존재한다. 본 논문에서는 이러한 문제를 해결하기 위해 Conditional Random Fields(CRF)를 이용하여 개체명과 문장목적을 동시에 인식하는 방법을 제안하며, 양질의 ETRI 개체명 사전을 이용해 전체적인 성능을 향상시켰다. 기존 시스템과의 비교 결과, 문장 단위 1.5%의 성능 향상을 보였고, 유의성 검정 결과 본 논문에서 제안하는 방법이 신뢰도 95%에서 통계적으로 유의하다는 결론을 얻었다.

키워드: 개체명 인식, Conditional Random Fields, 문장목적 인식, 대화시스템

Abstract A dialogue-based private secretary system provides some real-time information to a user using apps unlike a traditional dialogue system. The Spoken Language Understanding (SLU) module of the private secretary system consists of five components : domain, named-entity, object, operator and speech-act recognition. This paper proposes a simultaneous recognition method recognition of named entities and objects. The traditional dialogue-based private secretary system has some problems from using dictionary and rule-based resources. In order to solve these problems, the proposed system uses the conditional random fields (CRF) for named entity and object recognition. Moreover, the performance of the system has been improved by adding the high quality ETRI named-entity dictionary. In the empirical experiments, the proposed system results the higher performance of 1.5% in sentence level than a compared system using dictionary and rule based resources. In a significance test, we obtain a P-value of 0.11. This difference is considered to be statistically significant.

Keywords: named entity, conditional random fields, object, dialog system

· 본 연구는 산업자원통상부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]
· 이 논문은 제25회 한글 및 한국어 정보처리 학술대회에서 '대화형 개인 비서 시스템을 위한 하이브리드 방식의 개체명 및 문장목적 동시 인식기술'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 동아대학교 컴퓨터공학과
Blue772001@gmail.com
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
youngjoong.ko@gmail.com
(Corresponding author임)

논문접수 : 2013년 12월 19일
심사완료 : 2014년 2월 14일

정보과학회논문지: 소프트웨어 및 응용 제41권 제4호(2014.4)

1. 서 론

모바일 기술이 발전하면서, 우리는 언제, 어디서든 실시간으로 원하는 정보를 얻을 수 있게 되었다. 예를 들어 버스 정보를 실시간으로 얻을 수 있으며, 주식 정보 또한 실시간으로 조회가 가능해졌다. 사용자에게 정보를 제공하는 앱(App)의 종류가 다양해지면서 얻을 수 있는 정보 또한 기하 급수적으로 늘어났지만, 이런 많은 앱 중 양질의 앱을 선별, 통합하여 우리가 원하는 정보를 손쉽게 얻는 연구는 진행되지 않았다. 대화형 개인 비서 시스템은 사람에게 가장 편리한 인터페이스인 음성 인식을 통해 사용자가 원하는 정보를 파악하고, 기존의 여러 양질의 앱들을 통합, 활용하여 사용자에게 다양한 정보를 통합된 하나의 앱을 이용하여 제공해주는 시스템이다. 이러한 작업이 가능하기 위해선 사용자의 문장을 분석해 사용자가 원하는 정보가 무엇인지 파악하는 작업이 중요하다. 그 작업을 SLU(Spoken Language Understanding) 작업이라 하며 도메인, 개체명, 문장목적, 동작, 화행 인식으로 나누어진다. SLU 작업들 중 도메인 인식은 사용자의 질의에 따라 그에 맞는 앱을 실행시켜주는 인식 작업이므로 우선적으로 실행되어 도메인 오선택에 따른 SLU 전체 오류를 최소화 시켜야 하며, 동작과 화행 인식은 이전 발화 문장과 관련성이 있으며, 실마리 구간 탐색 방법을 사용할 수 없어 다른 접근 방법을 통해 인식 작업을 수행 해야한다. 그러므로 본 논문에서는 이러한 SLU 작업들 중 개체명 인식과 문장목적 인식을 동시에 수행하는 방법을 앞서 제안한다.

개체명 인식은 질의-응답 시스템과 정보검색 분야에서 유용하게 사용되고 있는 정보 추출의 한 단계이며, 개체명은 인명, 지명, 조직명, 시간, 날짜, 화폐 등의 고유명사이다. 개체명 인식은 문장에서 개체명을 식별하고 식별된 개체명의 종류를 결정하는 작업이며, 대화형 개인 비서 시스템에서는 문장에서 중요한 핵심어를 추출해 문장의 의미를 파악하는데 도움을 준다.

문장목적 인식은 사용자의 문장을 분석해 사용자가 원하는 정보가 무엇인지 찾아주는 인식 작업이다. 즉, 앱 구동을 통해 실행된 하나의 앱에서는 여러 정보가 제공되는데 사용자의 문장을 분석해 그 중 원하는 정보가 무엇인지 찾는것이다. 예를 들어, 날씨 앱을 구동 했을 경우 앱이 제공하는 정보인 강수량, 기온, 날씨 등의 정보들 중에서 사용자가 원하는 정보는 날씨 정보라는 것을 문장 분석을 통해 찾아주는 인식 작업이다.

대화형 개인 비서 시스템에서 개체명과 문장목적을 인식해 사용자가 원하는 정보를 파악하는 작업은 다음과 같은 프로세스로 진행된다.

입력 문장 : “오늘 기온에 대해 알려줘.”

인식된 개체명 : “오늘” 인식된 문장목적 : “기온”

대화형 개인 비서 시스템은 이 두 가지 인식 작업을 통해 개체명 “오늘”과 문장목적 “기온”을 추출하게 되고, 이 정보를 이용해 사용자가 원하는 정보는 “기온”에 대한 정보이며, 시기는 “오늘”이라는 것을 파악한다.

기존의 대화형 개인 비서 시스템은 사전-규칙 기반 방법을 사용하는데, 이 방법은 몇 가지 문제점을 가진다. 문제점으로는 데이터가 추가됨에 따라 지속적으로 규칙을 추가 해줘야 하는 문제, 규칙이 잘못 될 경우 많은 오류를 유발하는 문제, 파이프라인 방식으로 사전-규칙 기반을 적용함으로써 개체명 인식에서 생긴 오류로 인해 문장목적 인식에서의 규칙이 적용되지 않는 문제가 생기게 된다. 본 논문에서는 사전-규칙 기반의 문제점을 해결함과 동시에 개체명과 문장목적 인식이 서로 다른 인식 단위로 인해 각각 수행해야했던 기존의 SLU의 문제를 해결하는 방법을 제안한다. 즉, 실마리 구간 탐색 방법을 통해 두 인식의 인식 단위 문제를 해결하고, 통계 기반의 기계 학습 기법인 CRF와 양질의 ETRI 개체명 사전을 이용해 두 인식을 동시에 수행하는 방법을 제안하며, 기존 방법에 비해 성능을 향상됨을 보인다. 또한 5-fold cross validation을 수행하여 연구의 성과를 보여주며, 유의성 검정를 통해 제안하는 방법이 기존의 방법과 비교해 신뢰도 95%에서 통계적으로 유의함을 보인다.

본 논문은 다음과 같이 구성되어 있다. 1장의 서론에 이어 2장에서 관련 연구에 대해 살펴보고, 3장에서는 기존 방법과 문제점에 대해 설명하고, 4장에서는 제안하는 방법에 대해 설명하며, 5장에서는 실험 결과에 대해 기술하고, 6장에서는 결론 및 향후 연구과제에 대해 살펴본다.

2. 관련 연구

개체명 인식은 질의 응답 시스템과 정보 검색 분야에서 유용하게 사용되고 있는 정보 추출의 한 분야로서 문서나 문장 내에서 개체명을 추출하고 추출된 개체명의 종류를 식별하는 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 진행되어 왔으며, Conference on Computational Natural Language Learning 2002(CoNLL 2002)와 CoNLL 2003을 통해 더욱 많은 발전이 있었다[1]. 개체명 인식은 크게 3가지의 방법으로 연구되었다. 첫 번째는 규칙 기반 방법이며 이 방법에서는 주로 정규표현식[2]이나 자연어 특징을 이용한 규칙과 사전 정보[3]를 사용했다. 두 번째로 통계 기반

의 기계 학습 방법이며, 대표적인 방법으로는 Hidden Markov Model. Maximum Entropy Model, Conditional Random Fields, Decision Tree 등이 있다[4-7]. 마지막으로 규칙 기반과 기계 학습을 함께 사용한 하이브리드 방법도 연구되었다[8,9]. 이 방법은 규칙 기반과 기계 학습을 혼합함으로써, 규칙 기반, 기계 학습을 각각 수행한 방법보다 향상된 성능을 보였다.

문장목적 인식은 사용자 발화에서 질의의 목적이 무엇인지 파악해주는 작업이며, 대화 시스템에서 situation-based rules과 dialgue examples을 통합한 방법을 통해 문장목적을 인식하는 연구가 있었다[10]. 본 논문에서는 모바일 앱의 특징에 맞도록 문장목적 인식을 구축했다.

3. 기존 방법 : 개체명 및 문장목적 인식

이 장에서는 개체명과 문장목적을 인식하는 기존의 방법에 대해 기술하고, 문제점을 살펴본다.

기존의 대화형 개인 비서 시스템은 사전-규칙 기반 방법을 이용하며, 사전과 규칙을 혼합한 연구[11,12]와 비슷한 구조로 구성했다. 본 논문에서는 규칙을 정교하게 만들기 위해 정규표현식[2,11] 등 한국어 특징에 기반한 다양한 템플릿들을 만들어 사용했다. 또한 양질의 ETRI 개체명 사전과 함께 각 도메인별로 독립적인 규칙을 적용하여 오류를 최소화 하도록 구축했다. 그림 1은 사전정보와 규칙을 적용하는 방법을 보여준다.

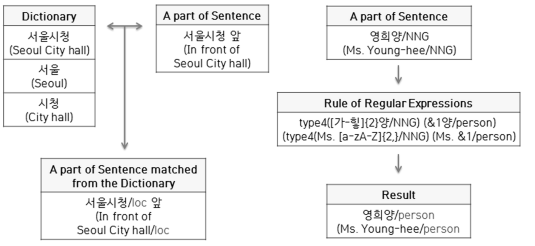

그림 1 사전 매치 방법(왼쪽)과 규칙 적용 방법(오른쪽)
Fig. 1 Dictionary-based method(Left) and rule-based method(Right)

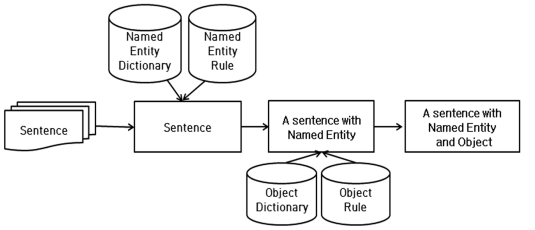

그림 2 기존의 사전-규칙 기반 방법의 시스템 구조도
Fig. 2 Architecture of traditional dictionary-rule based system

규칙은 여러 가지 유형에 유연하게 적용할 수 있도록 형태소 어휘, 형태소 정보, 정규표현식[2,11] 등 총 5가지의 규칙 템플릿을 적용했으며, 사전 매치는 최장 길이 일치법을 적용했다.

기존 시스템은 개체명과 문장목적 인식을 파이브라인 방식으로 수행해 개체명과 문장목적이 인식된 문장을 만들어냈다. 그림 2는 기존의 개체명과 문장목적을 인식하는 시스템의 구조도이다.

기존의 시스템은 크게 3가지 문제점을 가진다.

1. 데이터가 추가됨에 따라 규칙을 지속적으로 갱신해야 한다.
2. 규칙이 일반화되지 않을 경우, 심각한 오류를 유발한다.
3. 개체명, 문장목적 인식 작업을 파이프라인 방식으로 수행함으로써 문장목적 인식 단계에서의 규칙을 개체명이 인식된 문장을 기반으로 구성했기 때문에 개체명 인식 단계에서 인식 작업에 오류가 생길 경 우 문장목적 인식 단계에서 규칙이 적용되지 않아 오류가 생기는 2단계에 걸친 성능 저하가 발생한다.

우리는 이러한 문제점을 해결함과 동시에 기존의 방법과 비교해 인식 성능을 향상하는 방법을 제안한다.

4. 제안하는 방법 : 개체명 및 문장목적 동시 인식

이 장에서는 기존의 문제점을 해결하기 위한 방법으로 CRF를 이용해 개체명과 문장목적을 동시에 인식하는 방법을 제안하며, 양질의 ETRI 개체명 사전을 이용해 CRF의 성능을 개선시키는 방법을 보인다.

개체명 인식은 CRF와 양질의 ETRI 개체명 사전을 함께 활용함으로써 CRF를 단독으로 사용한 방법과 비교해 인식 성능을 향상시켰다. 그림 3은 사전 정보를 혼합한 CRF기반의 개체명 인식 시스템의 구조도이다.

그림 4에서 보는 바와 같이 문장에서 개체명 사전에 매치된 단어는 개체명 후보 단어로 인식하며 MUC7에서 사용한 BIO태그를 부착함으로써, CRF를 이용해 분류할 때 개체명 사전 자질로 이용할 수 있도록 구성했다.

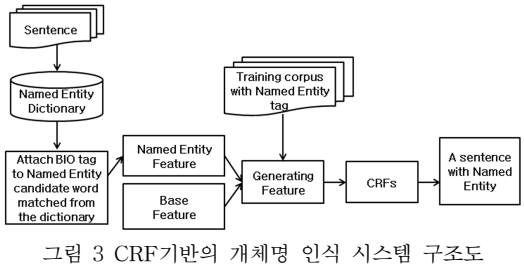

그림 3 CRF기반의 개체명 인식 시스템 구조도
Fig. 3 Architecture of CRF-based named entity recognition

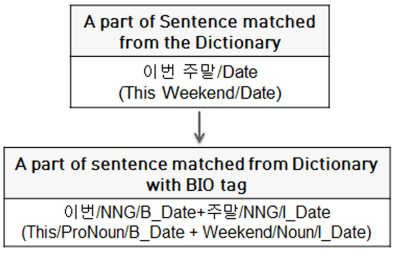

그림 4 BIO태그가 부착된 사전 정보

Fig. 4 Dictionary information with BIO tag

개체명과 문장목적 인식을 동시에 수행하기 위한 기반을 마련하기 위해서는 개체명과 문장목적의 인식 단위 문제를 해결해야 한다. 일반적인 개체명과 문장목적의 인식 단위는 개체명 인식-형태소 단위, 문장목적 인식-문장 단위이며, 문장 내에 개체명은 여러 개 존재할 수 있지만, 문장목적은 한 개만 존재한다는 차이점이 있다.

우리는 인식 단위 차이를 해결하기 위한 방법으로 문장목적 인식을 위한 실마리 구간 탐색 방법을 제안한다. 실마리 구간 탐색 방법은 문장 내에서 특정 문장목적으로 분류할 수 있는 핵심적인 구간을 선정하는 방법이다. 그림 5는 실마리 구간 탐색 방법의 예를 보여준다.

실마리 구간 탐색 방법을 통해 문장목적 인식 또한 개체명 인식과 같은 인식 단위를 사용해 인식 작업을 수행할 수 있으며, 개체명 사전과 달리 양질의 사전 정보가 따로 존재하지 않기 때문에 학습데이터를 이용해 사전을 구축했다. 그림 6은 사전 정보를 혼합한 CRF기반의 문장목적 인식 시스템 구조도이다.

그림 7에서 볼 수 있듯이 실마리 구간 탐색 방법을 통해 인식 단위의 차이를 해결함으로써, 개체명과 문장목적 인식 시스템이 같은 구조를 가지도록 만들었다.

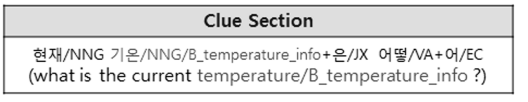

그림 5 실마리 구간 탐색 방법의 예

Fig. 5 Example of clue section search method

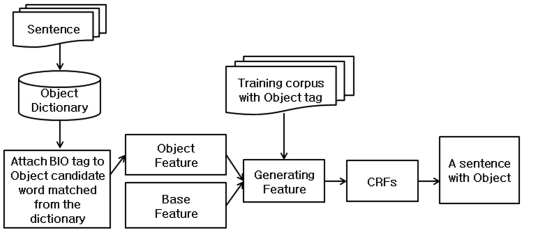

그림 6 CRF기반의 문장목적 인식 시스템 구조도

Fig. 6 Architecture of CRF-based object recognition system

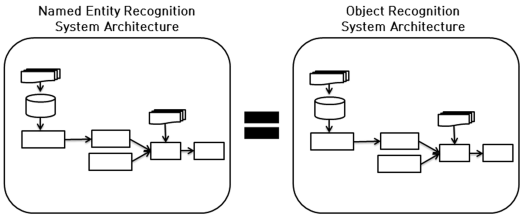

그림 7 개체명과 문장목적 인식 시스템 구조도

Fig. 7 Architecture of named entity and object recognition system

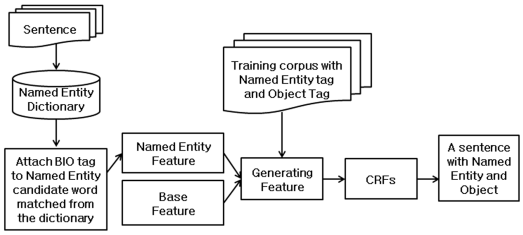

그림 8 개체명과 문장목적 동시 인식 시스템의 구조도

Fig. 8 Architecture of simultaneous recognition system for named entities and objects

두 인식 시스템의 구조를 같게 함으로써 인식을 동시에 수행할 수 있는 시스템 구축이 가능해진다. 그림 8은 본 논문에서 제안하는 개체명과 문장목적을 동시에 인식하는 시스템 구조도이다.

개체명과 문장목적을 동시에 인식하는 시스템을 구축할 때, 개체명과 문장목적 사전을 동시에 이용하여 개체명 및 문장목적 후보 단어를 추출 한 후, 개체명-문장목적 사전 자질을 각각 구축하지 않고, 개체명 사전만을 이용하여 후보를 추출해 개체명 사전 자질만을 구축한 이유는 3가지가 있다.

1. 문장목적 사전은 학습데이터를 이용해 구축한 사전이기 때문에 ETRI 개체명 사전과 달리 양질의 사전이 아니다.
2. 문장목적 인식 시스템에서 문장목적 사전 정보를 결합한 성능이 CRF만을 사용한 성능보다 낮았다.
3. 개체명 사전과 문장목적 사전을 동시에 이용해 후보 단어를 추출하여 자질을 각각 구축했을 때의 성능 이 개체명 사전만을 이용해 후보를 추출한 후, 자질 을 구축했을 때 보다 더 낮았다.

이러한 이유로 인해, 개체명과 문장목적을 동시에 인식하는 시스템에선 개체명 사전만을 이용해 후보를 추출하는 방법으로 시스템을 구축했다. 또한 동시 인식 시스템에서 한 형태소에 대해 개체명과 문장목적이 동시에 인식될 경우, [형태소 어휘/태그/BIO 개체명 태그/BIO 문장목적 태그]와 같은 템플릿을 사용해 한 형태

표 1 기본 자질
Table 1 base feature

morpheme lexicon/tag feature	1. The current morpheme lexicon / tag feature
morpheme lexicon sequence feature	2. Based on the current morpheme lexicon, lexicon feature located at (-2,-1,0,1,2)
morpheme tag sequence feature	3. Based on the current morpheme tag, tag feature located at (-2,-1,0,1,2)
feature in word	4. The current morpheme position in word 5. The current morpheme tag / word length

소에 중복으로 태그를 부착하는 방법으로 개체명 및 문장목적 태그 충돌 문제를 해결했다.

자질 집합은 형태소 어휘, 태그, 어절 내 자질 등 기본적으로 개체명 인식에 사용되는 자질 집합[5]과 인식 성능을 높이기 위해 본 논문에서 구축한 개체명 사전 자질을 사용했다.

표 1은 기본적인 자질을 보여준다.

기본 자질에 대한 부가 설명은 다음과 같다

1. 형태소 어휘/태그 자질 - 형태소 어휘/태그 정보
2. 형태소 어휘 시퀀스 자질
   2-1.w_0/w_1 2-2.w_{-1}/w_0 2-3.$w_{-1}/w_0/w_1$,
   2-4.$w_0/w_1/w_2$ 2-5.$w_{-2}/w_{-1}/w_0$
   2-6.$w_{-1}/w_0/w_1/w_2$ 2-7.$w_{-2}/w_{-1}/w_0/w_1$
   2-8.$w_{-2}/w_{-1}/w_0/w_1/w_2$
3. 형태소 태그 시퀀스 자질
   3-1.p_0/p_1 3-2.p_{-1}/p_0 3-3.$p_{-1}/p_0/p_1$
   3-4.$p_0/p_1/p_2$ 3-5.$p_{-2}/p_{-1}/p_0$
   3-6.$p_{-1}/p_0/p_1/p_2$ 3-7.$p_{-2}/p_{-1}/p_0/p_1$
   3-8.$p_{-2}/p_{-1}/p_0/p_1/p_2$

형태소 어휘/태그 시퀀스 자질은 현재 형태소 어휘/태그 (w_0,p_x)를 중심으로 이전(w_{-1}/w_0)/이후(p_0/p_1)의 연속된 형태소 어휘 및 태그 시퀀스 정보이며, -1은 현재 형태소 기준으로 바로 이전 형태소, -2는 현재 형태소 기준으로 2번째 앞의 형태소, 1은 현재 형태소 기준으로 바로 뒤의 형태소, 2는 현재 형태소 기준으로 2번째 뒤의 형태소를 나타냄

4. 형태소 어절 내 위치 자질-형태소가 어절의 시작, 중간, 끝 위치에 있는 지에 대한 정보
5. 형태소 태그/어절 길이 자질-형태소 태그와 어절 내 형태소의 개수 정보

표 2는 개체명 사전 자질을 보여준다.

표 2 개체명 사전 자질
Table 2 Named entity dictionary feature

Named entity dictionary feature	1. The current morpheme lexicon and Named entity information, matched from the NE dictionary
Pre-Named entity feature	2. Based on the current morpheme, all Pre-Named entity Information
Named entity presence feature	3. The current morpheme lexicon / tag and all named entity information in sentence
Named entity presence sequence feature	4. The current morpheme lexicon / tag and all named entity sequence information in sentence
Named entity sequence featrue	5. It is similar to morpheme tag sequence feature (-2,-1,0,1,2). If there is a morpheme matched from the NE dictionary, change morpheme tag->NE tag

개체명 사전 자질에 대한 부가 설명은 다음과 같다.

1. 개체명 사전 자질 - 사전에 매치된 개체명 후보 단어를 형태소 단위 별로 나누어 BIO 태그를 부착한 후, 현재 형태소를 기준으로 형태소 어휘 / BIO 개체명 조합 정보
2. 이전 개체명 자질 - 현재 위치 앞에 출현한 모든 BIO개체명 정보(NULL 정보 포함)
3. 개체명 존재 여부 자질 - 현재 형태소 어휘/태그 정보와 문장에 출현하는 모든 BIO 개체명 정보
4. 개체명 존재 여부 시퀀스 자질 - 현재 형태소 어휘/태그 정보와 문장에 출현하는 모든 BIO 개체명 시퀀스 정보(NULL 정보 포함)
5. 개체명 시퀀스 자질 - 형태소 태그 시퀀스 자질 $p_{-2}/p_{-1}/p_0/p_1/p_2$와 비슷하며, p_x위치의 형태소 태그에 BIO 개체명이 존재할 경우, p_x위치의 형태소 태그를 BIO 개체명 정보로 바꾸어 사용한 정보

5. 실 험

5.1 실험 대상 및 데이터 분포

본 논문에서는 6개의 도메인에서 8개의 개체명 태그셋과 28개의 문장목적 태그셋으로 이루어져 있으며, 1925개의 태깅된 데이터를 5-fold cross validation을 수행하여 평가했으며, 도메인별 데이터 분포는 그림 9와 같다.

모든 실험 결과는 정확도(accuracy)를 측정해 평가했으며, 개체명과 문장목적을 함께 인식하는 시스템은 문장 내의 모든 개체명과 문장목적이 정확히 일치하는 경우에만 정답으로 간주했으며, 기존 방법과 제안한 방법의 성능 차이가 통계적으로 유의한지 알아보기 위해 Student T-검정을 실시했다.

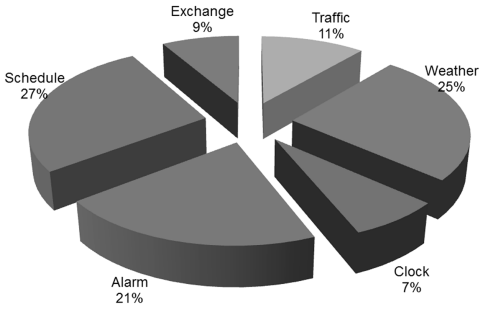

그림 9 도메인별 데이터 분포

Fig. 9 Domain-specific data distribution

5.2 실험 성능

실험은 개체명 인식, 문장목적 인식, 개체명과 문장목적을 함께 인식하는 방법 총 3가지로 나누어 성능을 평가했다.

개체명 인식의 성능은 그림 10과 같다.

개체명 인식의 성능은 ETRI 사전과 5가지의 유형 규칙을 적용한 사전-규칙 기반 방법의 시스템 성능이 높았으며, CRF와 ETRI 개체명 사전을 혼합한 방법이, CRF를 단독으로 사용했을 경우와 비교해 성능이 향상됨을 알 수 있었다.

문장목적 인식의 성능은 그림 11과 같다.

문장목적 인식은 CRF를 단독으로 사용 했을 경우 가장 높은 성능을 보이는 것을 확인할 수 있었다. 그 이유는 문장목적 인식은 구간을 통해 문장목적을 분류하기 때문에, 완전히 같은 단어 집합, 태그 정보가 반복으로 구간에 나타나는 경향(ex. 버스정류장, 기온, 추워, 달러를 원으로 등)이 많아 기본적인 자질이 더 높은 성능을 보이는 것으로 분석되었으며, 사전-규칙 기반 방법은 문장목적 특성상 규칙을 적용하기 까다로운 이유로 상대적으로 성능이 낮았다.

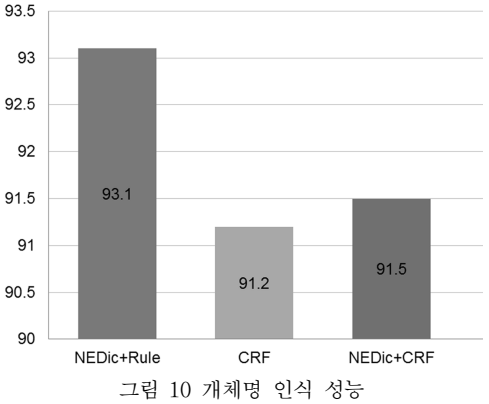

그림 10 개체명 인식 성능

Fig. 10 Performances of named entity recognition methods

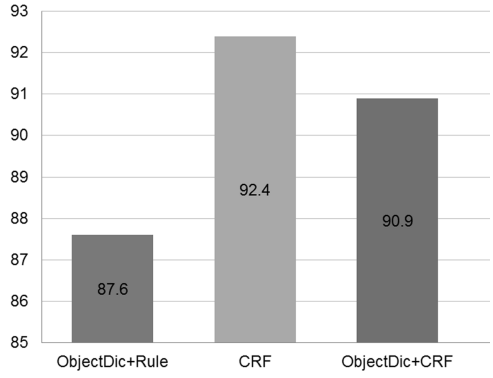

그림 11 문장목적 인식 성능

Fig. 11 Performances of object recognition

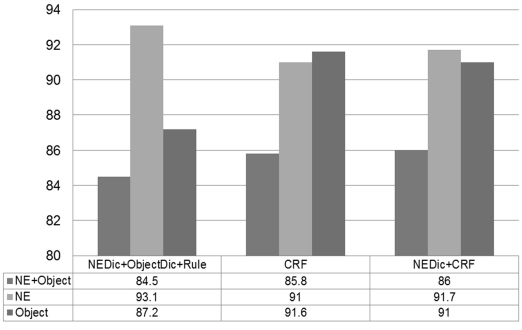

그림 12 개체명과 문장목적 인식 성능

Fig. 12 Performances of named entity and object recognition

개체명과 문장목적을 함께 인식하는 시스템의 성능은 그림 12와 같다.

사전-규칙 기반 방법은 개체명 단계에서 생긴 오류를 가지고, 문장목적 단계를 수행하기 때문에 개체명 인식에서 오류가 생길 경우, 문장목적에서 규칙이 적용되지 않아 오류가 축적될 수 있는 문제가 있었으며 우리가 제안하는 두 인식을 CRF를 통해 동시에 수행하는 방법이 기존의 사전-규칙 기반 방법과 비교해 높은 성능을 보였다. 또한 사전 정보를 CRF와 결합해 두 인식을 동시에 수행한 방법이 사전 정보를 사용하지 않았을 경우보다 약간의 성능이 향상되었다. 즉, 개체명 사전과 CRF를 혼합해 동시 인식한 시스템이 가장 높은 성능을 보였다.

사전-규칙 기반 방법과 본 논문에서 제안하는 개체명과 문장목적 동시 인식 시스템의 성능 차이가 통계적으로 유의한지 확인하기 위해 T-검정을 실시했다. 유의성 검정에는 도메인별로 유의성 검정[13]을 실시하는 방법과 5-fold 혹은 10-fold[14]로 유의성 검정을 실시하는 방법 등이 있다. 본 논문에서는 도메인별 유의성 검정보

표 3 T-검정 결과
Table 3 Result of one-sided paired T-test

P-value < 0.05(95%)	NEDic + CRF vs. NEDic+ObjectDic+Rule
One-sided Paired-T-Test	0.011

다 신뢰도가 높은 n-fold로 유의성 검정을 실시했다. 또한 데이터의 규모로 인해 10-fold를 수행할 경우 도메인별 테스트데이터가 작아져 도메인별 성능 평가에 문제가 생길 수 있는 이유로 5-fold로 실험을 평가했으며, 각 fold별 성능으로 유의성 검정을 실시했다. 표 3은 유의수준 0.05(신뢰도 95%)에서 유의성 검정의 결과이다.

T-검정에서 개체명 사전과 CRF를 결합한 방법의 성능 향상이 기존의 사전-규칙 기반 방법과 비교해 신뢰도 95%에서 통계적으로 유의미함을 확인할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 대화형 개인 비서 시스템의 SLU 작업들 중 개체명 인식과 문장목적 인식을 동시에 수행하는 방법을 제안 했으며, 양질의 ETRI 사전을 이용해 전체적인 성능을 개선시켰다. 6개의 도메인에서 8개의 개체명과 28개의 문장목적을 대상으로 개체명과 문장목적 분류를 수행한 결과, 기존의 사전-규칙 기반 방법보다 본 논문에서 제안한 방법이 문장단위 1.5%의 성능향상이 있었으며, T-검정결과 신뢰도 95%에서 통계적으로 유의함을 입증했다.

향후 연구로는 대화형 개인 비서 시스템의 SLU 작업 중 동작 인식과 화행 인식을 본 논문에서 제안하는 동시인식 시스템에 포함시켜, 도메인 인식을 제외한 모든 SLU 작업을 통합하는 방법에 대해 연구 할 것이다.

References

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

[2] 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.

[3] Kyung Hee Lee, Ju Ho Lee, Myung Seok Choi, Gil Chang Kim, "Study on Named Entity Recognition in Korean Text," *Proc of 12th Annual Conference on Human and Cognitive Language* pp.292-299, 2000.

[4] Nadeau, D. and Sekine, S., "A Survey of Named Entity Recognition and Classfication," *Lingvisicae Investigationes*, vol.30, no.1, pp.3-26, 2007.

[5] Changki Lee, Myeonggil Jang, "Named Entity Recognition with Structural SVMs and Pegasos algorithm," *Korean Journal of Cognitive Science 2010*, vol.21, no.4, pp.655-667.

[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] Cheongjae Lee, Sangkeun Jung, Jihyun Eun, Minwoo Jeong, Gary Geunbae Lee, "A situation-based dialogue management using Dialgogue examples," IEEE 2006.

[11] Khaled Shaalan, Hafsa Raza, "Person Named Entity Recognition for Arabic," ACL2007, Workshop, pp.17-24.

[12] Ali Elsebai, Farid Meziane, Fatma Zohra Belkredim "A Rule Based Persons Names Arabic Extraction System," Communications of the IBIMA Volume 11, 2009.

[13] Yiming Yang and Xin Liu, "A re-examinatio of text categorization methods," SIGIR'99.

[14] Janez Demsar, "Statistical Comparisons of Classifiers over Multiple Data sets," Journal of Machine Learning Research, 7(2006), 1-30.

이 창 수

2013년 동아대학교 컴퓨터공학과 학사 2013년~현재 동아대학교 컴퓨터공학과 석사과정. 관심분야는 자연어처리, 대화시스템, 기계학습 등

고 영 중

1996년 서강대학교 수학과 학사. 1996년~1997년 LG-EDS 근무. 2000년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년~현재 동아대학교 컴퓨터공학과 부교수. 관심분야는 자연어처리, 대화시스템, UI/UX, 빅데이터 분석, 텍스트마이닝(의견마이닝), 정보검색(교차언어검색) 등