열악한 환경의 음성 언어 이해를 위한 정보 추출 접근 방식*

은지현[1] 이창기[2] 이근배[1]
포항공과대학교 컴퓨터공학과[1]
한국전자통신연구원 미래기술연구본부 지식마이닝 연구팀[2]
{tigger,gblee}@postech.ac.kr[1]
leeck@etri.re.kr[2]

An Information Extraction Approach for Spoken Language Understanding in a Hostile Environment.

Jihyun Eun[1] Changki Lee[2] Gary Geunbae Lee[1]
Department of Computer Science and Engineering, POSTECH, Pohang, Korea[1]
Speech/Language Information Research Department,
Future Technology Research Division, ETRI, Daejeon, Korea[2]

요 약

본 논문에서는 환경 잡음과 원거리 음성 입력 그리고 노인 발화 등의 열악한 음성 인식 환경에서의 음성 언어 이해(spoken language understanding)를 위한 정보 추출 접근 방식에 대해 논하고 있다. 정보 추출의 목적은 미리 정의된 slot에 적절한 값을 찾는 것이다. 음성 언어 이해를 위한 정보 추출은 필수적인 요소만을 추출하는 것을 목적으로 하는 개념 집어내기(concept spotting) 접근 방식을 사용한다. 이러한 방식은 미리 정의된 개념 구조 slot에만 관심을 가지기 때문에, 음성 언어 이해에서 사용되는 정보 추출은 언어를 완전히 이해한다기보다는 부분적으로 이해하는 방식을 취하고 있다. 음성 입력 언어는 주로 열등한 인식 환경에서 이루어지기 때문에 많은 인식 오류를 가지고 이로 인해 텍스트 입력에 비해 이해하기 어렵다. 이러한 점을 고려하여, 특정 정보에 집중함으로써 음성 언어를 이해하고자 시도하였다. 도로 정보 안내 영역을 대상으로 한 실험에서 텍스트 입력(WER 0%)과 음성 입력(WER 39.0%)이 주어졌을 때, 개념 집어내기 방식의 F-measure 값은 각각 0.945, 0.823을 나타내었다.

1. 서 론

지금까지 음성 언어 이해를 위한 연구는 광범위하게 진행되어 왔다 [4][5][6][10][11]. 음성 문장의 구문론적 가능성과 의미론적 가능성을 설명하기 위해서 다양한 문법 규칙을 이용하는 언어 이해 체계는, 인간이 실제로 사용하는 매우 다양한 형태의 음성 문장이 주어지면 강건한 성질이 유지되기 어렵다. 더군다나 이러한 접근 방식은 원거리 음성 인식, 노인의 발화, 문법적으로 부정확한 문장, 그리고 잡음이 많은 환경일 경우 문장의 이해가 매우 어려워진다.

이러한 음성 인식의 한계를 극복하기 위해서 본 논문에서는 정보 추출(information extraction, IE) 기술을 바탕으로 필수적인 요소만 추출하는 것을 목적으로 하는 개념 집어내기(concept spotting) 접근 방식을 제안한다. 이러한 방식은 미리 정의된 개념 구조 slot에만 관심을 가지기 때문에, 언어를 완전히 이해한다기보다 부분적으로 이해하는 방식을 취하고 있다. 이 방식은 부분적인 언어 이해 방식이지만 특정 영역의 언어 이해를 위해 적절하게 설계된 slot이 있기 때문에 각 slot의 값으로부터 언어 이해에 필수적인 정보를 얻을 수 있다는 특징이 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 기존의 음성 언어 이해에 사용된 몇 가지 접근 방식을, 정보 추출 기반의 새로운 접근 방식과 비교하여 알아본다. 3절에서는 새로운 방법론을 자세하게 설명하고, 4절에서는 실험 결과와 분석을 싣는다. 마지막으로 5절에서는 결론 및 향후 연구 과제에 대하여 논한다.

* 본 연구는 과학기술부 뇌신경정보학 특정연구과제의 지원을 받아 수행되었음.

2. 관련연구

앞서 언급한 대로 다양한 문법 규칙을 이용하는 언어 이해 체계는 인간이 실제로 사용하는 매우 다양한 형태의 음성 문장이 주어지면 강건한 성질이 유지되기 어렵다. 이러한 단점이 나타나는 이유는 대부분 특정 영역에서 이루어지는 문장의 구문론적 분석이, 대부분 완전하지 않고 많은 오류와 애매성을 포함하고 있기 때문이다. 이러한 체계는 일반적으로, 수동으로 구축한 의미론적 단계의 문법과 강건한 파서(robust parser)를 통해 구현된다 [4][11]. 그러나 이러한 의미론적 문법을 통한 접근 방식은 고가의 개발비용이 요구되고, 사용자들이 특정 체계에서 지원되는 문법 규칙에 대해 무지할 경우 적절한 작업이 이루어지지 않을 가능성이 있다.

또 다른 접근 방식은 단어 문자열을 미리 정의된 의미 구조에 직접 할당하는 통계적 방법을 사용하는 것이다. 이 방식에서는 손수 조직된 문법이나 규칙은 학습 데이터로부터 자동으로 훈련되는 통계적 모델로 대체된다 [5][6][10]. 통계적 방법은 적절한 훈련 데이터가 주어지기만 하면 새로운 조건(새로운 작업 또는 새로운 언어)에서도 적절하게 적응될 수 있다는 점에서 매우 흥미롭다. 음성 언어 이해를 위한 통계적 방법은 이미 AT&T의 CHRONUS [6]와 BBN의 은닉 이해 모델 (Hidden Understanding Model, HUM) [10], Cambridge 대학의 HVS (Hidden Vector State) 모델 [5] 등을 통해 연구되어 왔다. 이런 모델은 주로 음성 인식과 자연 언어 처리에서 효과적인 성능을 보인 은닉 마코프 모델 (Hidden Markov Model, HMM)로부터 동기를 부여받았다.

HMM은 상태 천이 확률과 출력 확률 분포를 가지는 확률적 유한 상태 모델이다. 언어 처리 관련 분야에서 출력 확률은 일반적으로 불연속, 유한 단어 어휘에 대한 다항 분포로 표현된다. 그리고 Baum-Welch 알고리즘은 훈련 데이터에서 출력 열의 확률이 최대화 되도록 매개변수를 훈련하는데 사용된다. 전통적인 HMM 방식에는 크게 두 가지 문제점이 있다. 첫 번째 문제는, 예를 들면 전통적인 단어의 본질성에 추가로 대문자 사용, 단어 종결, 품사 정보와 같이 서로 비독립적이고 중복되는 다양한 자질들이 주어지듯이, 출력 값에 대한 표현이 많으면 많을수록 좋지만 이를 적절히 통합할 방법이 없다는 점이다. 두 번째 문제는, 전통적인 HMM 접근방식이 출력 열의 가능성을 최대화 시키도록 HMM 매개변수를 설정하고자 하지만, 대부분 언어 처리 응용에서는 주어진 출력 열로부터 상태 열을 예측하도록 이용되고 있다는 점이다. 다시 말하자면, 전통적인 접근방식의 관측과 관련된 조건부 문제를 해결하기 위해서 generative join model이 부적절하게 사용되고 있다는 점이다 [9]. 이러한 한계를 극복하고자 본 연구에서는 HVS 모델 [5]과 유사한 데이터 기반 방식을 최대 엔트로피 (Maximum Entropy, ME) 모델 [1] 하에서 풀어내고자 한다.

3. 정보 추출을 기반으로 하는 음성 언어 이해

본 논문에서는 음성 언어 이해 체계를 위해 개념 집어내기 접근 방식으로써 정보 추출 기술의 적용을 제안하였다. 이 장에서는 문장 단위의 개념 구조 slot 정의, 최대 엔트로피(Maximum Entropy, ME) 모델 [1]을 사용한 정보 추출 방법론, 그리고 시스템의 학습 및 평가 과정에 대해서 설명한다.

3.1 Slot 정의

예를 들어 도로 정보 안내와 같은 특정 영역에서 이루어지는 언어 이해에 정보 추출 기술을 적용하기 위해서 우선 몇 가지 slot을 정의해야 한다. 임의의 한 문장에는 서술어에 따른 문장 전체의 발화 의도를 나타내는 surface speech-act slot, 화자가 요구하는 직접적인 행동을 나타내는 action-type slot, 그리고 각 행동에 부가적인 여러 가지 component slot(예를 들면, 도메인에 따른 source, destination, via, address 등) 이 정의된다. 이 중 surface speech-act slot과 action-type slot을 대상으로 하는 정보 추출은 분류(classification) 문제로 취급된다. Surface speech-act slot의 값에는 yn-question, wh_question, request 등 여러 가지 문장 갈래 중 하나가 문장 단위로 할당된다. Surface speech-act slot과 비슷하게, action-type slot의 값은 특정 영역에서 등장하는 행동 중 한 가지가 할당된다. 도로 정보 안내 영역의 경우, confirm-route(길 확인), search-route(길 탐색), search-address(주소 탐색) 등이 이에 해당한다. Source(출발지), destination(목적지), via(경로), address (주소) 등 action component slot에 대한 정보 추출은 개체 명 인식(named entity recognition, NER) 문제로 취급되고, 각각의 개체 명은 구문론적 구(句) 단위로 인식된다.

3.2 Maximum Entropy Classification Model

최대 엔트로피(Maximum Entropy, ME) 모델 [1]은 주어진 제약 조건을 만족하는 여러 확률 분포 중에서 가장 균일한 분포 상태를 가지는 모델이다. 바꾸어 말하면, ME 모델은 주어진 제약 조건 하에서 최대 엔트

로피를 가지는 확률 분포를 가지고 있다. 이를 수식으로 나타내면 아래와 같다.

P = { models consistent with constraints }

$H(p)$ = Entropy of $p, p \in P$

$$P_{ME} = argmax_{p \in P} H(p) \qquad (1)$$

여기서 P_{ME}가 최대 엔트로피 확률 분포를 가지는 모델이다.

특히, ME 모델은 다양한 이질적인 정보를 통합하는 데 유용한 구조를 가지고 있다. ME 모델의 목적은 주어진conditional probability를 최대화시키는 y값을 찾는 데 있다. 정보 추출 기반의 음성 언어 이해를 위한 ME 모델에서, x는 문장의 문맥이고, y는 특정 영역에서 문장을 상대로 미리 정의된 slot의 값이다. k개의 자질 제약조건이 주어졌을 때, conditional probability는 다음과 같다.

$$p(y|x) = \frac{1}{Z(x)} exp\left(\sum_{i=1}^{k} \lambda_i f_i(x, y)\right) \qquad (2)$$

여기서,

k는 자질 개수,

f_i는 각 자질,

λ_i는 각 자질의 가중치 매개변수,

$Z(x)$는 $\sum_y p(y|x) = 1$을 보장하기 위한 정규화 요소이다.

ME 모델의 가장 두드러진 특징은 모델의 특성을 완전히 드러내는 후보 자질들을 선택해 주기만 하면 된다는 점이다. 그 외의 일은 ME 모델 알고리즘에 따라 작업이 진행된다. 여러 후보 자질 중에서 ME 모델로 표현하는데 있어 의미 있는 자질을 선택하는 일과 선택된 자질의 가중치 매개변수를 계산하는 일은 ME 모델 알고리즘에서 해결된다.

본 논문에서는 후보 자질로서 구문론적 자질과 의존관계 자질을 사용하였다. 구문론적 자질은 품사 부착(part-of-speech tagging)과 구 묶음(phrase chunking)과 같이 여러 구문 분석 단계를 거쳐서 나온 결과로 이루어진다. 의존 관계 자질은 의존 관계 파싱 (dependency relation parsing)을 통한 주요어구-수식어구의 관계 분석 결과로 이루어진다.

ME 모델의 매개변수 추정에 사용되는 알고리즘에는 Generalized Iterative Scaling (GIS) [3], Improved Iterative Scaling (IIS) [2], 그리고 Limited Memory BFGS (L-BFGS) [7] 등 잘 알려진 것이 몇 가지 있다. 본 논문에서는 대규모 비선형 최적화를 위한 L-BFGS 알고리즘을 사용하였다.

3.3 학습 및 평가 과정

본 논문에서 제시한 정보 추출 기반의 음성 언어 이해 체계에서는 ME 모델을 학습하기 위해서 텍스트 문장을 사용하였다. 학습에 사용된 자질은 문장의 구문/의존 관계 분석 결과로부터 선택됐고, 미리 정의된 구조의 slot-value 쌍, 즉 정답은 음성 언어 말뭉치를 semantic frame tagging한 결과로부터 얻을 수 있다. 그러면 매개변수 추정 후에 특정 모델의 매개변수를 얻을 수 있다. 훈련된 모델의 매개변수는 특정 slot-value가 짝지어질 가능성을 나타내기 때문에 음성 언어 형태의 평가 데이터가 주어지면 slot 값의 확률 분포를 계산할 수 있다. 그리고 이 분포에서 가장 높은 확률 값을 가지는 것을 미리 정의된 slot의 정답으로 선택한다. ME 모델을 이용한 학습 및 평가 과정은 Fig 1.와 같다.

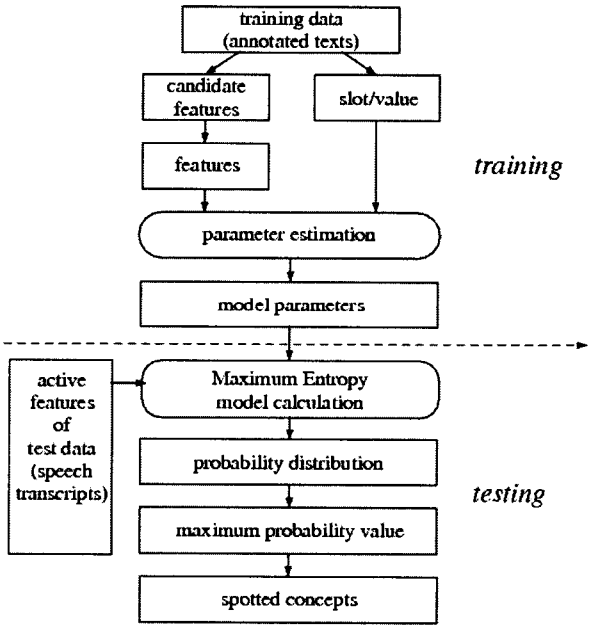

Fig 1. 학습 및 평가 과정

4. 실 험

본 실험에서는 정보 추출 기반의 한국어 음성 언어 이해 체계를 구현하였고, 학습 데이터로써 도로 정보 안내 영역의 말뭉치를 사용하여 다양한 실험을 수행했다.

4.1 실험 설정

도로 정보 안내 영역 말뭉치는 semantic frame tagging 된 462개 한국어 문장으로 이루어져 있고, 이

말뭉치에는 텍스트 문장과 음성 문장(음성 인식기 결과) 두 가지 유형이 있다. 한국어 음성 인식기는 LG전자기술원에서 만든 것을 사용하였고, 도로 정보 안내 영역에서 음성 인식 성능은 WER(word error rate) 39.0%이다. 본 실험에서 정보 추출 기반의 음성 이해 시스템은 텍스트 ;ᄆᆞᆫ잔ᄋᆞ ᄀᆞᆫ학습됐고, 텍스트 문장과 음성 문장으로 평가 결과를 비교하였다.

본 체계의 성능을 평가하기 위해서 각 문장은 surface speech-act slot, action-type slot, 여러 component slot에 대해 semantic frame tagging이 요구된다. Surface speech-act와 action-type은 한 문장 전체로부터 결정되기 때문에, 이 두 slot은 문장 단위로 산출된 자질들을 사용하였다. 반면에 action component slot은 한 문장 내에서 각 구문론적 구(句) 단위로 결정된다. 그래서 이 경우 현재 구 뿐만이 아니라 문맥 정보를 위해 이전, 이후 구로부터 추출된 자질까지 사용하였다. 예를 들어 한 문장의 slot-value 구성은 아래와 같다.

여기서	잠실까지	가는데	한남대교를	거치나요?
source	destination	-	via	-

surface = yn_question
action = confirm_route
source = 여기
destination = 잠실
via = 한남대교
→ component slots

본 실험의 평가 척도로는 일반적으로 잘 알려진 재현률(recall), 정확률(precision), 그리고 이것들의 조화평균인 F-measure를 이용하였다.

$$F-measure = \frac{2PR}{P+R} \quad (3)$$

4.2 실험 결과

본 논문에서의 실험은 크게 두 가지로 나눌 수 있다. 첫 번째 실험은 정보 추출 기반의 음성 언어 체계가 음성 입력에 대해서 얼마나 강건함을 유지하는지에 대한 것이다. 두 번째 실험은 자질 선정에 있어 의존 관계 자질이 음성 언어 이해 향상에 도움이 되는지에 대한 것이다.

4.2.1 텍스트 입력과 음성 입력일 때 성능 비교

본 실험에서는 도로 정보 안내 영역의 462개 한국어 문장으로 10-fold cross validation을 수행하였다. 텍스트 문장으로 의미 구조를 추출한 결과가 음성 문장을 사용한 결과보다 월등히 뛰어나다는 것을 Table 1,2.에서 확인할 수 있다. 여기서 텍스트 문장은 WER 0%인 음성 문장으로 간주될 수 있다. 주목할 점은 WER 39.0%인 음성 문장이 주어졌을 때, 본 논문에서 제안한 정보 추출 기반의 음성 언어 이해 성능에서 보이는 두 입력 간의 성능 차이는 F-measure 0.122 에 불과하다는 사실이다. 이러한 결과는 정보 추출 구조를 기반으로 하는 개념 집어내기 방식이, 보다 높은 WER을 가지는 음성 입력에 대해서도 강건함을 유지한다는 것을 보여준다.

	재현률(R)	정확률(P)	F-measure
surface	0.989	0.989	0.989
action	0.993	0.993	0.993
comp. slots	0.888	0.937	0.912
micro average	0.960	0.931	0.945

Table 1. 텍스트 입력(WER 0%)에 대한 결과

	재현률(R)	정확률(P)	F-measure
surface	0.870	0.870	0.870
action	0.896	0.896	0.896
comp. slots	0.732	0.834	0.779
micro average	0.855	0.793	0.823

Table 2. 음성 입력(WER 39.0%)에 대한 결과

4.2.2 의존 관계 자질을 사용했을 때 성능 비교

두 번째 실험은 자질 선정에 있어 의존 관계 자질이 음성 언어 이해에 얼마나 기여하는지에 대한 것이다. 비교 대상은 구문론적 자질(syntactic features, SF)만을 사용한 경우와 추가적으로 의존 관계 자질(dependency relation features, DRF)을 사용한 경우이다. 구문론적 자질에는 품사 정보, 구 단위 정보, 어휘 정보가 있다. 의존 관계 자질은 의존 관계 파싱으로 주요어구-수식어구의 관계를 분석한 결과에서 현재구의 주요어구에 대한 정보로 이루어진다. 한 문장에서 나타나는 의존 관계를 Fig 2.에서 살펴볼 수 있다.

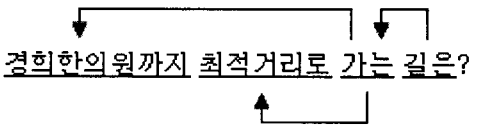

Fig 2. 의존 관계 예제

Table 3.에서 보다시피, WER 39.0% 인 음성 입력이 주어졌을 때 구문론적 자질만 사용한 경우와 추가적으로 의존 관계 자질을 사용한 경우, 각각 F-measure 0.779, 0.780이다. 비록 근소한 차이지만 의존 관계 자질이 음성 언어 이해 향상에 도움을 준다는 것을 알 수

있다.

의존 관계 자질을 사용한 실험은 구 단위로 인식되는 component slot을 대상으로 했다.

		재현률(R)	정확률(P)	F-measure
텍스트 입력	SF	0.888	0.937	0.912
	+DRF	0.890	0.938	0.913
음성 입력	SF	0.732	0.834	0.779
	+DRF	0.730	0.838	0.780

Table 3. 의존 관계 자질 사용에 따른 결과

5. 결 론

본 논문에서는 열악한 음성 인식 환경에서 다양하고 비독립적인 자질을 사용하여, 음성 언어 이해를 위한 새로운 방법으로써 정보 추출 접근 방식을 제안하였다. 또한 정보 추출 기반의 음성 언어 이해를 위한 패턴 분류 방법으로는 Maximum Entropy 모델을 적용하였다. 그리고 실험 결과에서 알 수 있듯이, 문장에서 다양한 관찰 자질을 추출하여 미리 정의된 의미 구조 slot의 값을 찾는 방법을 이용한 정보 추출 기술이 음성 언어 이해에 효과적으로 작용한다는 사실을 확인했다.

여러 실험에서 보다시피, 한국어 음성 인식기의 성능은 WER 39.0%로 인식률이 꽤 낮은 음성 문장이 입력으로 주어졌지만 텍스트 입력과 음성 입력의 언어 이해 성능 차이는 F-measure 0.122에 불과하였다. 따라서 도로 정보 안내 영역에서의 성능 평가 결과는 본 논문에서 제안한 정보 추출 기반의 음성 언어 이해 시스템이 많은 오류를 포함한 음성 언어가 입력으로 주어져도 강건함을 유지한다는 것을 보여준다. 또한 의존 관계 파싱의 결과로부터 새로운 의존 관계를 추출하여 자질로써 사용했을 때, F-measure 0.001 정도로 성능 향상에 도움을 준다는 것을 확인하였다.

현재 본 연구는 한국어로 된 도로 정보 안내 영역에 적용되고 있으나, 추후에는 다른 여러 음성 언어 이해 체계와 체계적인 비교를 하기 위해 ARPAR ATIS (Air Travel Information System) 말뭉치를 사용해 영어 쪽으로 확장해 볼 계획이다.

참고 문헌

[1] Berger, A., Della Pietra, S. and Della Pietra, V., "A maximum entropy approach to natural language processing", Computational Linguistics, 22(1):39-71, 1996.

[2] Berger, A., "The Improved Iterative Scaling Algorithm: A Gentle Introduction", School of Computer Science Carnegie Mellon University December, 1997.

[3] Darroch, J. and Ratchli, D., "Generalized Iterative Scaling for Log-Linear Models", The Annals of Mathematical Statistics, Vol. 43, No. 5, pp. 1470-1480, 1972.

[4] Dowding, J., Moore, R., Andry, F., and Moran, D., "Interleaving syntax and semantics in an efficient bottom-up parser", in Proc. of ACL, 1994, pp. 110-116.

[5] He, Y. and Young, S., "A Data-Driven Spoken Language Understanding System", IEEE Workshop on Automatic Speech Recognition and Understanding, US Virgin Islands, 2003.

[6] Levin, E. and Pieraccini, R., "CHRONUS, the next generation", in Proc. of the DARPA Speech and Natural Language Workshop, 1995, pp. 269-271.

[7] Liu, D. C. and Nocedal J., "On the Limited Memory BFGS Method for Large Scale Optimization", Math. Programming, 1989.

[8] Manning, C. D. and Schutze H., Foundations of Statistical Natural language Processing, MIT Press. Cambridge, 1999.

[9] Mccallum, A., Freitag, D., and Pereira, F., "Maximum Entropy Markov Models for Information Extraction and Segmentation", in Proc. ICML, 2000.

[10] Miller, S., Bates, M., Bobrow, R. Ingria, R., Makhoul, J., and Schwartz, R., "Recent progress in hidden unders - tanding models", in Proc. of the DARPA Speech and Natural Language Workshop, 1995, pp. 276-280.

[11] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications", Computational Linguistics, 1992, 18:1, pp. 61-86.