베이지안 모형 기반 한국어 의미역 유도

원유성○, 이우철, 김형준, 이연수
㈜엔씨소프트
{styner, darkgeo, hjunk, yeonsoo}@ncsoft.com

Bayesian Model based Korean Semantic Role Induction

Yousung Won○, Woochul Lee, Hyungjun Kim, Yeonsoo Lee
NCSOFT Corp.

요 약

의미역은 자연어 문장의 서술어와 관련된 논항의 역할을 설명하는 것으로, 주어진 서술어에 대한 논항 인식(Argument Identification) 및 분류(Argument Labeling)의 과정을 거쳐 의미역 결정(Semantic Role Labeling)이 이루어진다. 이를 위해서는 격틀 사전을 이용한 방법이나 말뭉치를 이용한 지도 학습(Supervised Learning) 방법이 주를 이루고 있다. 이때, 격틀 사전 또는 의미역 주석 정보가 부착된 말뭉치를 구축하는 것은 필수적이지만, 이러한 노력을 최소화하기 위해 본 논문에서는 비모수적 베이지안 모델(Nonparametric Bayesian Model)[1][2]을 기반으로 서술어에 가능한 의미역을 추론하는 비지도 학습(Unsupervised Learning)을 수행한다.

주제어: 비모수적 베이지안 모형, 비지도 학습, 의미역 결정

1. 서론

의미역이란, 문장 내에서 서술어에 의해 기술되는 행위나 사태와 관련된 논항의 역할을 뜻한다. 문장에서 나타난 서술어와 논항의 의미역이 무엇인지 결정하는 일은 문장이 나타내고 있는 의미를 파악하기 위한 중요한 단계로써 형태소 분석, 개체명 인식, 의존 구조 분석과 더불어 자연언어처리 분야에서 상당히 중요한 기반 기술이라고 할 수 있다. 이는 기계 번역, 정보 추출, 질의 응답과 같은 다양한 자연언어처리 영역에서 활용되는 만큼 많은 연구가 이루어지고 있다[3][4]. 대표적으로 FrameNet[5]과 Propbank[6] 말뭉치를 기반의 기계 학습을 통한 의미역 결정 시스템이 주를 이루고 있지만, 이러한 말뭉치를 구축하는 것이 상당히 어렵기 때문에 본 논문에서는 의미역 부착 말뭉치의 도움 없이 문장 내에 나타난 서술어와 논항 사이의 관계에 대한 특성을 비모수적 베이지안 모형을 통한 비지도 학습을 통하여 각 서술어가 가질 수 있을 법한 의미역을 유도하는 것을 목표로 한다. 2장에서는 현재까지 의미역 결정 연구에 대한 간략히 소개하고, 3장에서는 베이지안 모형에 관한 이론적 배경과 더불어 의미역 유도를 위한 모델링 방법을 설명하고, 4장에서는 비지도 학습 방법으로 유도된 의미역에 대한 평가가 이루어지며, 마지막으로 5장에서는 결론에 대해 기술한다.

2. 관련 연구

의미역 결정에 관한 연구에는, 전통적으로 서술어가 요구하는 통사적인 논항 정보를 가진 격틀 사전에 기반한 선택 제약(Selectional Restriction) 방법[7]과 의미역 주석이 부착된 말뭉치에 이용한 기계 학습 방법이 있다.

격틀 사전 기반 의미역 결정은 주어진 문장의 서술어-논항 관계를 격틀과의 유사도에 따라 사전에 명시된 의미역을 결정하는 것으로 높은 정확률을 보이지만, 수많은 경우에 따른 서술어와 논항들 사이의 격틀 사전을 구축하는 것은 불가능에 가까울 수 있다. 이에 대한 대안으로, 말뭉치에 포함된 의미역 주석 정보를 학습데이터로 이용한 기계 학습 방법을 통해 다양한 경우에 대한 처리를 가능하게 하였다.

한국어 의미역 결정을 위한 대표적인 연구로 Korean Propbank를 학습하는 Support Vector Machine을 이용[3]한 방법과 최근 널리 활용되는 Deep Neural Network를 이용[4]한 지도 학습 방법이 있다. 두 방법 모두 상당히 높은 성능의 의미역 결정 시스템을 보여주고 있지만, 현실적으로 의미역에 대한 주석 정보를 포함한 다량의 학습데이터를 이용할 수 없는 환경인 동시에, 앞선 방법과 마찬가지로 양질의 학습데이터를 구축하는 것 또한 상당히 많은 시간과 비용을 수반하게 되는 단점이 있다.

영어권에서는 이러한 한계를 극복하기 위한 비지도 학습 방법으로, 베이지안 모형을 고안하여 서술어와 논항 사이에서 발견할 수 있는 자질(Feature)에 대한 군집(Cluster)을 발견하였고, 이를 통해 서술어 별로 존재할 수 있는 의미역을 유도하였다[8][9]. 하지만 의미역을 대변할 수 있는 자질을 단순하게 사용하였고, 추론에 의한 자질의 군집에 의미역 레이블을 부여하는 시도를 하지 않아 실제 의미역 결정 시스템에서 추론된 군집이 어떠한 역할을 할 수 있을지에 대한 분석을 하지 않았다.

본 논문에서는 베이지안 모형을 그래프 모형과 함께 수식으로 구체화하고 한국어 의미역을 위한 자질을 반영하였다. 이를 통해 한국어 서술어에서 존재 할 수 있는 의미역을 비지도 방식으로 유도하고 다양한 실험 조건에 따른 양상과 함께 유도된 의미역에 자동으로 레이블을

부여하여 입력 문장에 대한 의미역 결정이 이루어질 수 있음을 보인다.

3. 접근 방법

3.1 문제 정의

의미역 결정을 위한 일반적인 과정은 다음과 같다.

1. 서술어 인식 및 분류 (Predicate Identification / Classification)
2. 논항 인식 및 분류 (Argument Identification / Labeling)

서술어 인식 및 분류 단계는, 문장 속에서 서술어를 찾아내고 해당 서술어가 여러 가지 의미를 가질 경우 어떤 의미인지를 구분하는 역할을 한다. 논항 인식 및 분류 단계는 앞서 확인된 서술어와 관련된 논항을 찾아내고 해당 논항의 역할이 무엇인지 파악한다. 본 논문의 연구 범위는 논항 분류 단계에 그 목적이 있다. 다시 말하면, 의미역 주석 정보 없이 비지도 학습을 통하여 각 서술어 별 의미역을 유도하는 것이 본 연구의 주요 목표이므로, 이후부터 기술되는 연구 내용은 문장 속에서 서술어와 관련된 논항이 무엇인지 주어진 상황으로 한정한다.

3.2 비모수적 베이지안 모형

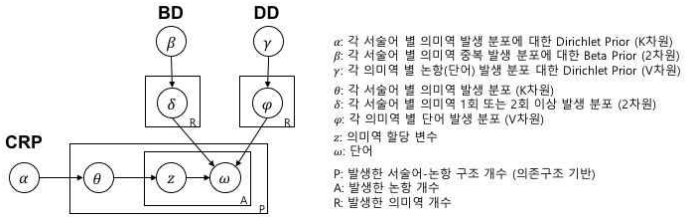
본 논문에서 다루는 베이지안 모형은 베이지안 확률론(Bayesian Probability)에 기반한 것으로, 어떤 가정에 대한 믿음, 즉 사전확률분포(Prior Distribution)가 존재하고 관측된 데이터에 의해 사후확률분포(Posterior Distribution)를 도출하여, 이것을 새로운 믿음(사전확률분포)으로 갱신하는 철학을 담고 있다. 다시 말하면, 사전확률분포, 즉 모수(Parameter)는 알려져 있지 않은 확률변수(Random Variable)이며, 여기서 비모수적 베이지안 모형이란, 모수가 무한한 차원(Infinite Dimensional Space)에서 정의되는 것을 의미한다. 이러한 이론적 배경을 기반으로 다음과 같은 가정을 통해 각 서술어의 의미역 유도를 위한 베이지안 모형을 구성한다.

1. 임의의 문장에서 특정 서술어는 논항과 관련된 의미역의 분포가 존재: 문장 내에 있는 각 서술어와 논항 사이에는 의미역을 나타내는 자질(Argument Key)의 군집이 존재한다.
2. 임의의 서술어에서 특정 의미역의 중복 발생 가능성에 대한 분포가 존재: 특정 의미역은 서술어와 결부하여 1번 또는 여러번(2번 이상) 발생하는 경우가 존재한다.
3. 임의의 의미역 내에 발생 가능한 단어의 분포가 존재: 각 의미역을 구성하는 단어의 군집이 존재한다.

이것은 토픽 모형(Topic Model)의 대표적인 예인 잠재 디리클레 할당(Latent Dirichlet Allocation)[10]을 의미역 결정 분야에 응용한 것으로 볼 수 있고, 군집의 개수가 정해져 있지 않은 보다 더 일반적인 관점에서의 베이지안 모형으로 생각 할 수 있다.

3.3 그래프 모형(Graphical Model)

의미역 유도를 위해 베이지안 확률을 구성한 그래프 모형은 다음과 같다.



[그림 1] 의미역 유도를 위한 그래프 모형
CRP: Chinese Restaurant Process
BD: Beta Distribution
DD: Dirichlet Distribution

[그림 1]은 3.2에서 언급한 세 가지 가정을 다항 분포 및 이항분포(Multinomial Distribution / Binomial Distribution)와 이것의 켤레사전확률분포(Conjugate Prior Distribution)인 디리클레 분포 및 베타 분포(Dirichlet Distribution / Beta Distribution)와의 혼합 모형(Mixture Model)으로 구성한다.

이 모형에 따른 첫 번째 가정은, 학습데이터에 존재하는 문장이 가진 서술어-논항 구조는 다음과 같이 Θ에 의한 의미역 분포를 갖도록 한다.

$$(\Theta_1, \Theta_2, \ldots, \Theta_R) \sim CRP(\alpha)$$

이때 의미역의 개수를 나타내는 Θ의 차원(R)은 디리클레 과정(Dirichlet Process)을 설명하는 대표적인 방법 중의 하나인 Chinese Restaurant Process를 통해 무한대의 차원으로 모델링하여 R값을 고정적인 값이 아닌 각 서술어에 맞는 의미역의 개수(R)를 알아낼 수 있도록 한다.

두 번째 가정으로, 각 서술어에서 나타나는 특정 의미역은 다음과 같이 δ에 의해 의미역이 1번만 나타날지 여러번(2번 이상) 나타날지에 대한 분포를 디리클레 분포의 특수한 형태인 2차원의 베타 분포로 설정 한다.

$$(\delta_1, \delta_2) \sim Beta(\beta)$$

세 번째 가정인, 각 의미역 내에 발생 가능한 단어 분포는 다음과 같이 ϕ에 의한 단어 분포를 V차원의 디리클레 분포로 설정한다.

$$(\phi_1, \phi_2, \ldots, \phi_V) \sim \mathrm{Dir}(\gamma)$$

위와 같이 베이지안 모형이 구성되었을 때, 본 연구의 목적은, 초모수(α, β, γ)가 주어진 상황에서 데이터를 관측하였을 때, 모수(Θ, δ, ϕ)의 발생 가능성, 즉 사후확률이 최대인 모수를 찾는 것이다. 이것은 베이즈 정리(Bayes' Theorem)에 의해 다음의 결합 확률을 최대로 하는 것으로 생각할 수 있다.

$$P(w,z,\theta,\delta,\phi|\alpha,\beta,\gamma)$$
$$=\prod_{i=1}^{R}P(\phi_i|\gamma)\prod_{i=1}^{R}P(\delta_i|\beta)\prod_{j=1}^{P}P(\theta_j|\alpha)\prod_{j=1}^{P}\prod_{k=1}^{A}P(z_{j,k}|\theta_j)\prod_{j=1}^{P}\prod_{k=1}^{A}P(w_{j,k}|\phi_{z_{j,k}},\delta_{z_{j,k}})$$
$$=\prod_{i=1}^{R}P(\phi_i|\gamma)\prod_{i=1}^{R}P(\delta_i|\beta)\prod_{j=1}^{P}P(\theta_j|\alpha)\prod_{j=1}^{P}\prod_{k=1}^{A}P(z_{j,k}|\theta_j)\prod_{j=1}^{P}\prod_{k=1}^{A}\frac{P(\phi_{z_{j,k}}|w_{j,k})P(\delta_{z_{j,k}}|w_{j,k})P(w_{j,k})}{P(\phi_{z_{j,k}})P(\delta_{z_{j,k}})}$$

여기서 식을 Θ, δ, ϕ에 대해 주변화(Marginalize) 함으로써 관측하는 문장들마다 가질 수 있는 변동성을 줄이는 목적과 더불어 모형을 단순화 할 수 있다. 따라서 위 식을 적분하고,

$$P(w,z|\alpha,\beta,\gamma)=\int_\theta\int_\delta\int_\phi P(w,z,\theta,\delta,\phi|\alpha,\beta,\gamma)d\theta d\delta d\phi$$
$$=\int_\theta\prod_{j=1}^{P}P(\theta_j|\alpha)\prod_{j=1}^{P}\prod_{k=1}^{A}P(z_{j,k}|\theta_j)d\theta\cdot\int_\delta\prod_{i=1}^{R}P(\delta_i|\beta)\prod_{j=1}^{P}\prod_{k=1}^{A}P(w_{j,k}|\delta_{z_{j,k}})d\delta\cdot$$
$$\int_\phi\prod_{i=1}^{R}P(\phi_i|\gamma)\prod_{j=1}^{P'}\prod_{k=1}^{A}P(w_{j,k}|\phi_{z_{j,k}})d\phi\cdot\prod_{j=1}^{P'}\prod_{k=1}^{A}\frac{1}{P(w_{j,k})}$$

마지막 상수항을 제외한 뒤, 감마함수를 이용하여 다시 정리하면 다음과 같은 식으로 유도할 수 있다.

$$\prod_{j=1}^{P}\left(\frac{\Gamma(\sum_{i=1}^{R}\alpha_i)\prod_{i=1}^{R}\Gamma(n^i_{j,(\cdot)}+\alpha_i)}{\prod_{i=1}^{R}\Gamma(\alpha_i)\Gamma(\sum_{i=1}^{R}n^i_{j,(\cdot)}+\alpha_i)}\right)\prod_{i=1}^{R}\left(\frac{\Gamma(\sum_{d=1}^{2}\beta_d)\prod_{d=1}^{2}\Gamma(n^i_{(\cdot),d}+\beta_d)}{\prod_{d=1}^{2}\Gamma(\beta_d)\Gamma(\sum_{d=1}^{2}n^i_{(\cdot),d}+\beta_d)}\right)\prod_{i=1}^{R}\left(\frac{\Gamma(\sum_{v=1}^{V}\gamma_v)\prod_{v=1}^{V}\Gamma(n^i_{(\cdot),v}+\gamma_v)}{\prod_{v=1}^{V}\Gamma(\gamma_v)\Gamma(\sum_{v=1}^{V}n^i_{(\cdot),v}+\gamma_v)}\right)$$

위에서 나타난 모수 Θ, δ, ϕ는 모두 초기 설정한 초모수(Hyperparameter)에 따라 뒤에서 설명할 베이지안 추론(Bayesian Inference) 과정을 통해 데이터를 가장 잘 설명할 수 있도록, 모형의 전체 확률값이 가장 높은 방향으로 결정 된다.

3.4 베이지안 추론(Bayesian Inference)

위의 식에서 볼 수 있듯이, 앞서 구성한 베이지안 모형은, 결국 각 서술어마다 논항으로써 역할을 하는 단어(ω)가 어떤 의미역(군집)에 할당(z)되는지에 따라 위의 확률값을 계산할 수 있고, 관측한 데이터를 가장 잘 설명하는 모수 Θ, δ, ϕ을 도출 할 수 있다. 이때, 가능도(Likelihood)로 사용한 다항분포의 켤레사전확률분포가 디리클레 분포이기 때문에 혼합 모델에서의 사후확률의 계산이 간단해지게 된다.

구체적으로 설명하면 디리클레 프로세스(DP)에서 샘플링한 분포(G)는 다음과 같이 디리클레 분포(Dir)로 표현할 수 있고,

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$$

α : concentrate parameter

H : base distribution

A : partition of observation

이것의 사후분포 또한 아래와 같이 디리클레 분포를 만족하게 된다.

$$(G(A_1), \ldots, G(A_r))|x_1, \ldots, x_r \sim \mathrm{Dir}(\alpha H(A_1)+n_1, \ldots, \alpha H(A_r)+n_r)$$

즉, 위에서 나타난 n 값에 따라 [그림 2]와 같은 분할(Partition)을 형성하며 사후확률의 갱신이 이루어지게 된다.

다시 본론으로 돌아가면, 관측된 단어가 어떤 군집에 해당하는지를 찾아내는 것이 목적인데, 본 연구에서는 추론 과정의 복잡도를 줄이고 문장 내에서 발견할 수 있는 자질을 의미역에 최대한 반영하기 위하여 관측된 단어 기준이 아닌 관측된 자질(Argument Key)을 기준으로 의미역을 할당하도록 하였다.

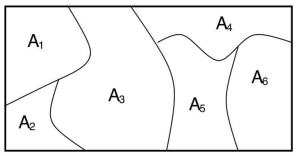
본 논문에서 사용한 Argument Key의 구성[8][9][1]은 다음과 같다.

1. VOICE: ACT/PASS (능동/수동)
2. POSITION: LEFT/RIGHT(서술어에 대한 논항의 상대적 위치)
3. POS: 논항의 품사 태그
4. DEP: 의존 관계 레이블
5. NE: 논항의 개체명
6. ENDING: 논항의 조사 정보
7. PATH: 서술어-논항 사이의 의존 구문 트리 경로

Argument Key는 문장 내에서 발견할 수 있는 서술어와 논항 사이의 자질로써 [그림 2]와 같이 의미역 집합의 분할(Partition), 즉 집합족(family of sets)을 구성하

1) 영어권 연구[8][9]에서는 1, 2, 4의 자질을 기반으로 한 군집(의미역)을 만들어 실험함, 본 연구에서 실험 결과 자질을 많이 사용할수록 군집의 Purity는 높아지지만 Collocation은 낮아지는 경향을 보임(자질의 복잡도가 높아지므로 생성된 군집의 개수가 많아지기 때문)

게 된다.



[그림 2] 의미역 집합의 분할(Argument Key Partition)

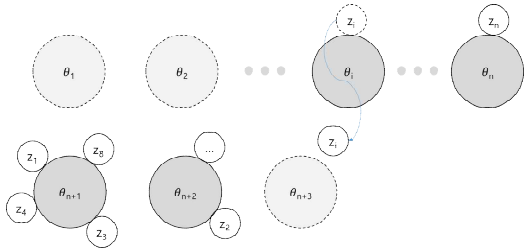
이것은 Argument Key의 의미역(군집) 할당 상태에 따라 형성되는 집합족에 속하는 하나의 경우라고 볼 수 있는데, 이것은 다음에 나오는 [표 1]에서와 같이 잠재변수(Latent Variable) z에 의해 결정 된다.

Argument Key	z (Cluster ID)
ACT:LEFT:SW:SBJ:XN:은:NP-SBJ	1
ACT:RIGHT:VA:XDE:XN:XE:VP-_-VP-MOD-NP-_	1
ACT:LEFT:VCN:CMP:XN:XE:VP-CMP	1
...	...
ACT:LEFT:NNG:AJT:XN:에:NP-AJT-VP-_	2
ACT:RIGHT:NNG:SBJ:PS:은:VP-MOD	2
ACT:LEFT:NNG:AJT:XN:만큼:NP-AJT	2
...	...
ACT:LEFT:NNG:OBJ:AM:을:NP-OBJ	3
ACT:LEFT:SL:SBJ:OGG:는:NP-SBJ-VP-_-VP-_-VP-_	3
ACT:LEFT:SL:SBJ:OGG:도:NP-SBJ	3
...	...
ACT:LEFT:NNG:OBJ:XN:은:NP-OBJ-VP-_	4
ACT:LEFT:NNG:SBJ:PS:만:NP-SBJ	4
ACT:RIGHT:VX:AJT:XN:고:VP-_-VP-_	4
...	...
ACT:LEFT:NR:OBJ:QT:를:NP-OBJ	5
ACT:LEFT:NR:OBJ:XN:XE:NP-OBJ	5
ACT:LEFT:NR:SBJ:QT:은:NP-SBJ	5
...	...

[표 1] Argument Key의 할당

베이지안 추론은 위의 [표 1]에서처럼 z가 어떤 값을 가질 수 있는지를 하나씩 탐색해나가면서 앞서 기술한 확률값을 계산하게 되는데, 직관적으로도 복잡도가 굉장히 크다는 것을 알 수 있다. 즉, Aargument Key의 개수를 x라 하고 군집의 개수를 y라 하면 y^x만큼의 경로가 존재하고, 비모수적 베이지안 모형에 대한 추론 과정의 경우에는 군집의 개수 y가 고정된 것이 아니기 때문에 탐색범위가 무한에 가깝다고도 볼 수 있다. 따라서 본 연구에서는 Greedy한 탐색 방법을 적용하였다.

먼저 초기 설정으로 각 Argument Key에 각기 다른 군집(의미역)를 할당하고, 데이터를 가장 잘 설명할 수 있는 방향으로 각 군집을 병합해 나간다.



[그림 3] Greedy Search 방식의 베이지안 추론 과정

이때, Argument Key의 발생 빈도가 큰 순서대로 군집을 재할당하도록 하여 많이 관측된 자질이 먼저 군집화(Clustering)가 이루어지도록 한다. 위 [그림 3]으로부터 앞서 언급한 Chinese Restaurant Process를 대략적으로 이해할 수 있다. 예를 들어 손님(Argument Key)이 식당에 들어와서 테이블(의미역, 군집)을 차지하는데, 사람들이 많이 앉아 있는 테이블에 앉을 확률과 새로운 테이블에 혼자 앉을 확률을 비교하여 어디에 앉을 지를 결정하는 것으로 생각 할 수 있다. 이러한 과정을 통해 디리클레 프로세스를 설명할 수 있고, 추론 과정 또한 비슷한 방식으로 이해할 수 있다. 이것은 Markov Chain Monte Carlo(MCMC) 알고리즘 중의 하나인 Gibbs Sampling[11]으로도 설명되는 것으로, 변수를 하나씩 연쇄적으로 회를 반복하여 갱신하는 방식으로 앞서 설명한 Argument Key의 집합족을 탐색할 수 있게 된다.

4. 실험 방법 및 결과

4.1 학습데이터 및 평가데이터

비지도 학습에 의한 한국어 의미역 유도를 위해서는 문장에 나타난 서술어와 논항이 무엇인지에 대한 정보가 주어져야 한다. 이를 위해 본 연구에서는, 먼저 신뢰성 있는 의미역 결정 시스템을 이용하여 야구 기사 텍스트를 분석하였다. 이때, 의미역 결정 시스템이 판단한 서술어와 의미역 레이블을 지닌 개체를 논항으로 간주하였고, 여기서 의미역 레이블을 제외한 서술어-논항 정보와 해당 문장을 학습데이터로 사용하였고, 이중 일부는 레이블까지 포함하여 평가데이터로 사용하였다.

4.2 실험 방법

본 연구에서 설명한 베이지안 모형은 확률적인 추론 과정을 통해 의미역 주석 정보 데이터의 도움 없이 각 서술어와 관련 있는 논항과 내재하고 있는 자질 정보의 군집화를 이루어 의미역을 유도하게 되고, 각 의미역마다 그럴듯한 단어 군집을 만들어 내는 데 그 목적이 있다. 따라서 실험은 각 서술어마다 발생한 논항의 군집화[2]가 얼마나 잘되었는지(의미역 유도)가 주된 목표이고, 추가적으로는 이렇게 형성된 군집이 어떤 의미역 레이블과 관련되어, 실제 의미역 결정이 가능한 지를 분석하는 것이다. 서술어 별로 유도된 의미역의 군집화가 잘 되었는지를 측정하기 위해 Purity와 Collocation을 측정하였고, 각 군집에 의미역 레이블을 할당하였을 때, 평가 문장에 대한 의미역 결정 결과를 Precision과 Recall

2) 본 실험에서는 야구 기사에서 많이 쓰이는 서술어 중에서 "(경기가) 열리다", "(홈런을) 치다"에 대한 부분 실험을 수행

로 측정하였다.

Purity(PU)는 서술어 마다 추론된 각각의 군집을 기준으로 정답 군집(본 실험에서는 의미역 결정 시스템에 의한 군집) 사이에 논항을 공유하는 정도를 측정하고, Collocation(CO)은 반대로 정답 군집을 기준으로 추론된 군집과의 논항을 공유하는 정도를 측정한다.

$$PU = \frac{1}{N}\sum_{i}\max_{j}|G_j \cap C_i|$$

$$CO = \frac{1}{N}\sum_{j}\max_{i}|G_j \cap C_i|$$

G: 정답 군집
C: 추론된 군집
N: 전체 논항 수

앞서 언급한 것처럼 추가적인 실험으로써 유도된 군집이 어떤 의미역을 나타내는지 레이블을 부여하는 것이 필요한데, 본 연구에서는 임의의 추론된 군집 내에 존재하는 Argument Key와 단어 분포를 후보 정답 군집 내에 존재하는 Argument Key와 단어 분포와의 유사성을 Cross-Entropy로 측정하여, 그중 가장 작은 값을 가지는 후보 정답 군집의 의미역 레이블을 할당하는 방식을 적용하였다.

$$H(p,q) = -\sum_{x} p(x)\log q(x)$$

x: Argument Key 또는 논항

4.3 실험 결과

실험은 학습데이터 사이즈와 Greedy Search에서의 반복 회수(Iteration)에 따른 군집화 성능과 의미역 결정 성능의 추이를 측정하였다.

열리							
Iteration	# of Cluster	Clustering			Labeling		
		Purity	Collocation	F1	Precision	Recall	F1
1	21	0.7489	0.4241	0.5415	0.8419	0.8178	0.8297
2	21	0.7922	0.5120	0.6220	0.8230	0.7994	0.8110
3	21	0.7865	0.5095	0.6184	0.8427	0.8186	0.8305
4	21	0.7951	0.5117	0.6227	0.8310	0.8072	0.8189
5	21	0.7669	0.5085	0.6115	0.8112	0.7880	0.7994
6	20	0.7826	0.5128	0.6196	0.8352	0.8113	0.8230
7	20	0.7885	0.5071	0.6173	0.8415	0.8174	0.8293
8	20	0.7970	0.5052	0.6184	0.8440	0.8199	0.8317
9	20	0.7904	0.5922	0.6771	0.8448	0.8207	0.8326
10	21	0.7930	0.5090	0.6200	0.8465	0.8223	0.8342

치							
Iteration	# of Cluster	Clustering			Labeling		
		Purity	Collocation	F1	Precision	Recall	F1
1	44	0.6507	0.3203	0.4293	0.7701	0.7353	0.7523
2	44	0.6447	0.3261	0.4331	0.7102	0.6782	0.6938
3	44	0.6456	0.4200	0.5089	0.7102	0.6782	0.6938
4	44	0.6511	0.3286	0.4368	0.7165	0.6842	0.7000
5	44	0.6470	0.4026	0.4963	0.7543	0.7203	0.7369
6	44	0.6362	0.3182	0.4242	0.6772	0.6466	0.6615
7	44	0.6463	0.4179	0.5076	0.6992	0.6677	0.6831
8	44	0.6559	0.3383	0.4464	0.6819	0.6511	0.6662
9	44	0.6557	0.4307	0.5199	0.6882	0.6571	0.6723
10	44	0.6453	0.4074	0.4995	0.7024	0.6707	0.6862

[표 2] 서술어 “열리”, “치” 에 대한 의미역 유도 결과
열리: 문장 3932개 (서술어-논항 구조 10545개) 학습
치: 문장3888개 (서술어-논항 구조 8906개) 학습
Labeling P, R, F1: 의미역 레이블 별 Micro Average

Greedy Search에서는 회를 반복하면서 앞서 3.4절에서 언급한 Argument Key의 Partition을 구성하게 되는데 [표 2]에서 볼 수 있듯이, 매 회마다 군집화 성능 및 각 군집에 의미역 레이블을 부여했을 때 의미역 결정 성능의 변화가 있다는 것을 볼 수 있다. 이것은 각 회에서 Argument Key가 어느 의미역에 할당 되었는지에 따라 결정되는데, 이상적인 경우 회를 반복할수록 더 나은 분포를 만들어 내고 결과적으로 좋은 군집화를 이루게 된다.

본 실험 결과에서는 서술어 “열리”는 반복 회수가 9일 때, “치”는 반복 회수가 5일 때 가장 좋은 의미역의 군집화 및 레이블 부여가 이루어졌다. 이것은 실험에서 사용한 서술어-논항 구조 정보가 의미역 결정 시스템에서 자동적으로 생성해 낸 것을 이용한 점과, 문장에서 찾을 수 있는 자질 정보(Argument Key의 생성)의 불완전성, 그리고 베이지안 모형에 초기값을 설정되는 초모수[3] 값에 따라 많은 영향을 받으므로 여러 조건하에서 추가 실험이 필요하다.

열리		
Cluster ID	Argument	Label
1	새천년홀	ARGM-LOC
1	'	
1	타이중	
1	1라운드	
1	CC	
1	목동구장	
1	SK	
1	'	
1	KIA	
1	대구구장	
...	...	...
2	홈경기	ARG1
2	결승전	
2	원정경기	
2	개막전	
2	올스타전	
2	아시아선수권	
2	월드베이스볼클래식	
2	준결승전	
2	넥센전	
...	...	...

치		
Cluster ID	Argument	Label
2	홈런	ARG1
2	안타	
2	적시	
...	...	...
3	LG	ARG0
3	SK	
3	KIA	
3	가슴	
...	...	...
15	이범호	ARG0
15	갈	
15	강봉규	
15	로페즈	
...	...	...
36	5월	ARGM-TMP
36	2010년	
36	15일	
36	19일	
36	2일	
36	20일	
...	...	...

[표 3] 서술어 “열리”, “치”의 의미역 유도 결과

위 [표 3]에서는 각 서술어 별로 유도된 의미역과 해당 의미역 내에 군집화 된 단어 분포를 볼 수 있다. 여기서 Cluster ID는 베이지안 추론 과정에서 형성된 군집을 뜻하고 [표 2]에서 볼 수 있듯이 생성된 군집의 수(# of Cluster)는 서술어 마다 다르고 스스로 결정된다. 또한 표의 오른쪽 열에 각 군집마다 부여된 레이블은 앞서 설명한 정답 후보 군집과의 Cross-Entropy 값을 통해 정답 군집의 Arguemnt Key와 단어 분포와의 유사성이 높은 레이블을 할당한 결과이다[4]. 위의 결과로 미루어 보아 서술어 별로 형성된 자질의 군집이 어느 정도 의미역을 표현 할 수 있다고 생각 할 수 있다. 다만 군집화 성능이

3) 실험에서 사용한 초모수(Hyperparameter) 값
α=1000, β_1=1, β_2=0.1, γ=0.00000001

4) 이 부분은 어느 정도의 정답 군집의 Arguemnt Key와 단어 분포가 필요하고, 본 실험에서는 Argument Key 분포 : 단어 분포 = 2 : 8 로 설정하여 각각의 Cross-Entropy 값을 가중합 하여 사용함

레이블 부착 성능과 반드시 비례하지 않는 것을 볼 수 있는데, 이것은 향후 Arguemnt Key 생성의 고도화를 이루거나, 자동으로 인식한 논항 인식의 결과의 노이즈를 줄인다면 더 나은 결과를 보일 것이라 기대한다.

열리						
Label	Total	Retrieved	Matched	Precision	Recall	F1
ARG0	98	0	0	-	-	-
ARG1	547	800	498	0.6225	0.9104	0.7394
ARG2	20	0	0	-	-	-
ARGM-ADV	1	0	0	-	-	-
ARGM-CND	2	0	0	-	-	-
ARGM-DIS	14	0	0	-	-	-
ARGM-EXT	1	0	0	-	-	-
ARGM-INS	8	0	0	-	-	-
ARGM-LOC	902	789	758	0.9607	0.8404	0.8965
ARGM-MNR	3	0	0	-	-	-
ARGM-PRD	1	0	0	-	-	-
ARGM-TMP	841	789	753	0.9544	0.8954	0.9239
AUX	10	0	0	-	-	-
Average	2448	2378	2009	0.8448	0.8207	0.8326

치						
Label	Total	Retrieved	Matched	Precision	Recall	F1
ARG0	146	103	97	0.9417	0.6644	0.7791
ARG1	270	344	250	0.7267	0.9259	0.8143
ARG2	19	1	0	-	-	-
ARG3	2	0	0	-	-	-
ARGM-ADV	4	0	0	-	-	-
ARGM-CAU	9	13	3	0.2308	0.3333	0.2727
ARGM-CND	3	0	0	-	-	-
ARGM-DIS	14	2	2	1.0000	0.1429	0.2500
ARGM-EXT	18	38	12	0.3158	0.6667	0.4286
ARGM-INS	2	0	0	-	-	-
ARGM-LOC	36	36	28	0.7778	0.7778	0.7778
ARGM-MNR	32	6	6	1.0000	0.1875	0.3158
ARGM-PRP	1	0	0	-	-	-
ARGM-TMP	67	54	44	0.8148	0.6567	0.7273
AUX	42	38	37	0.9737	0.8810	0.9250
Average	249	188	132	0.7021	0.5301	0.6041

[표 4] 비지도 학습에 의해 형성된 자질의 군집에 레이블을 부여했을 때의 의미역 결정 성능 (각 레이블에 대한Micro Average Precision)
Total: 평가 대상 서술어-논항 개수
Retrived:비지도 학습 기반 의미역 유도 + 의미역 레이블 부착 기반으로 의미역 결정이 이루어진 논항 개수
Matched: 의미역 결정 결과가 정답과 일치하는 논항 개수

5. 결론

본 논문에서는 한국어 의미역 결정 문제를 지도 학습에 의한 접근 방법이 아닌 베이지안 모형에 기반한 비지도 학습을 통해 의미역 레이블 부착 말뭉치 없이, 각 서술어 별로 존재할 법한 의미역을 유도하고 실제 의미역 결정 문제까지 확장되었을 때 상당히 높은 성능의 의미역 결정이 가능하다는 것을 보였다. 대부분의 자연언어처리와 관련되어있는 문제들은 모호성이 크기 때문에 주석 정보가 포함된 말뭉치를 구축하는 작업이 굉장히 어려운 경우가 많다. 따라서 해당 문제를 풀 수 있는 직관적인 가정과 믿음을 반영한 베이지안 모형을 구성하고 데이터를 잘 설명할 수 있는 모수를 추론해 낸다면 의미역 결정 문제 처럼 수동으로 논항의 역할을 구분하기 힘든 부분을 추상적으로 표현 할 수 있고, 향후 적절한 레이블을 통해 그 역할을 구체화 할 수 있을 것이다.

향후 연구로는, 먼저 의미역 결정 시스템에 의한 불완전한 학습데이터가 아닌 Korean Propbank와 같은 양질의 말뭉치를 활용하여 동일한 실험을 재현해 보고, 초모수 값에 따른 실험 결과의 양상을 분석해 볼 계획이다. 또한 베이지안 추론 과정에서의 방대한 탐색 범위를 좀더 효율적으로 찾아가는 방법을 고안해야 할 것이다. 이는 결국 추론 과정의 속도를 높여 다양한 환경에서의 실험을 가능하게 하고 더 많은 서술어의 의미역 유도를 위한 필수적인 요소라고 할 수 있다. 마지막으로 의미역 결정 문제에서 선행되어야 하는 서술어 인식 및 분류, 논항 인식과 관련한 비지도 학습방법에 대한 연구를 통해 의미역 유도 및 결정을 위한 전 과정을 비지도 학습 방법을 통해 의미역 결정 시스템을 구축해 볼 계획이다.

참고문헌

[1] Gershman and D. Blei. A tutorial on Bayesian nonparametric models. *Journal of MathematicalPsychology*, 56:1-12, 2012.

[2] P. Orbanz and Y.W. Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*, pages 81-89, 2010.

[3] 이창기, 임수종, 김현기. Structural SVM 기반의 한국어 의미역 결정. 정보과학회논문지 42.2 220-226, 2015.

[4] 배장성, 이창기, 임수종. 딥 러닝을 이용한 한국어 의미역 결정. 한국정보과학회 학술발표논문집 690-692, 2015.

[5] Baker, Collin F., Charles J. Fillmore, and John B. Lowe, The berkeley framenet project, Proc. of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998.

[6] Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon, Korean Propbank, [Online]. Available: http://catalog.ldc.upenn.edu/LDC2006T03

[7] 김병수, 이용훈, 이종혁, 비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정, 정보과학회논문지, Vol. 34, No. 2, 2007.

[8] Lang, Joel, and Mirella Lapata. "Unsupervised semantic role induction via split-merge clustering." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[9] Titov, Ivan, and Alexandre Klementiev. A Bayesian approach to unsupervised semantic role induction. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[10] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. Journal of machine Learning research, 993-1022, 2003

[11] Andrieu, Christophe, et al. An introduction to MCMC for machine learning. Machine learning 5-43, 2003.