질의 응답 시스템을 위한 반교사 기반의 정답 유형 분류

박선영O, 이동현, 김용희, 류성한, 이근배
포항공과대학교, 컴퓨터공학과
{sypark322, semko, ttti07, ryush, gblee}@postech.ac.kr

Semi-Supervised Answer Type Classification For Question-Answering System

Seonyeong ParkO, Donghyeon Lee, Yonghee Kim, Seonghan Ryu, Gary Geunbae Lee
Department of Computer Science and Engineering, POSTECH

요 약

기존 연구에서는 질의 응답 시스템에서 정답 유형을 분류하기 위해 패턴 매칭 방식이나 교사 학습(Supervised Learning)을 이용했다. 패턴 매칭 방식은 질의 분석을 통해 수동으로 패턴을 구축해야 한다. 교사 학습에서는 훈련 데이터 전체에 정답 유형이 태깅(Tagging)되어야 하며, 이를 위해서는 사용자의 질의에 정답 유형을 수동으로 태깅하는 작업이 많이 필요하다. 웹을 통해 정답 유형이 태깅되지 않은 대용량의 사용자 질의 말뭉치를 구할 수 있지만, 이 데이터에는 정답 유형이 태깅되어 있지 않다. 따라서, 대용량의 사용자 질의에 비례하여, 정답 유형을 수동으로 태깅하는 작업량이 증가한다. 앞서 언급한 두 가지 방법론에서, 정답 유형 분류를 위해 수작업이 많이 필요하다는 문제점을 해결하고자 본 논문에서는 일부 태깅된 훈련 데이터를 필요로 하는 반교사 학습(Semi-supervised Learning)에 기반한 정답 유형 분류를 제안한다. 이는 정답 유형 분류 작업에 필요한 노동력을 최소화함으로 대용량의 데이터를 통한 효율적 질의 응답 시스템 구축을 가능하게 한다.

주제어: 정답 유형, 질의 응답 시스템, 잠재 디리쉴레 할당(Latent Dirichlet Allocation, LDA)

1. 서론

질의 응답 시스템은 많은 양의 정보를 바탕으로 사용자의 질문에 정확한 답을 찾아주는 시스템이다. 질의 응답 시스템은 기존의 검색엔진과 다르게, 불필요한 정보를 제외하고, 사용자가 찾고자 하는 정보만을 제공한다는 장점이 있다. 따라서, 질의 응답 시스템이 제공하는 서비스는 정보 검색의 궁극적인 목표와 부합하며, 빅 데이터 시대에 정보의 효율적 사용이 필요하다는 측면에서 각광받고 있다. 이러한, 질의 응답 시스템 개발은 해외에서 뿐만 아니라 국내에서도 중요한 이슈로 떠오르고 있다.

질의 응답 시스템은 크게 질의 분석 단계, 정답과 관련된 문서 추출 단계, 문서로부터 정답을 추출하는 단계로 나눠져 있다. 질의 분석 단계에서는 정답과 관련된 문서검색을 위하여, 사용자의 질의에서 키워드, 정답 유형 등을 추출한다. 질의 응답 시스템에서 정답 유형이란 사용자가 질의를 통해 찾고자 하는 정답의 유형을 말한다. 따라서, 질의 분석 단계에서 분석된 정답 유형은 문서에서 정답을 찾을 때, 검색 제약 조건으로 활용된다.

사용자 질의에서 정답 유형을 추출하는 작업은 질의 응답 시스템에 필수적인 요소이다. 대부분의 질의 응답 시스템 개발에서 정답 유형 정보를 활용하고 있다[1-8]. 정답 유형 결정은 n가지 정답 유형들 중 하나의 유형으로 분류하는 문제로 정의 되어 왔다. 기존의 시스템 개발에서는 사용자의 질의에서 정답 유형을 결정하기 위해 교사학습을 이용하거나 패턴과 규칙을 활용했다[1-8]. 표 1은 정답 유형과 질의에 대한 예를 보여준다.

표 1 정답 유형의 예

정답 유형	질의
Animal	What is the Ohio state bird?
Human	Who wrote "the divine Comedy"?
City	What is the capital of Mongolia?
...	...

패턴과 규칙에 기반한 방법을 사용하기 위해서는 사용자 질의를 분석하여 수작업으로 패턴을 구축해야 한다. 최근의 질의 응답 시스템에서는 패턴과 규칙만을 이용하여 정답 유형 분류를 하지 않고, 통계 모델을 함께 사용하는 경우가 대부분이다[2-6,8]. 통계 모델을 사용하는 최근의 정답 유형 분류 연구는 교사 학습을 이용한다[2~6,8]. 교사 학습에서 정답 유형 분류 모델 훈련을 위해서는 각 사용자 질의에 정답 유형이 모두 태깅된 훈련 데이터가 필요하다. 사용자 질의 데이터는 웹을 통해서 쉽게 수집할 수 있지만, 사용자 질의에 정답 유형이 태깅된 데이터는 구하기 어렵다. 따라서 기존 연구에서는 사용자 질의에 정답 유형을 수동으로 태깅하여 훈련 데이터를 제작했다[2-6,8].

정답 유형 수동 태깅에 대한 작업량을 줄이고자, 본 논문에서는 교사 학습 방법 대신에 반교사 학습 방법을 이용하여 정답 유형을 분류하였다. 반교사 학습은 일부만 태깅된 훈련 데이터를 통해 정답 유형 분류 모델 제작이 가능하다. 본 논문에서는 반교사 학습 기반의 잠재 디리쉴레 할당(Semi-Supervised Latent Dirichlet

Allocation, Semi-Supervised LDA)을 이용하였다. 일부 태깅된 데이터를 이용하여 정답 유형을 분류하였고, 정확도를 측정하였다. 또한 같은 양의 태깅된 훈련 데이터를 이용하였을 때 교사 학습 방법에 의한 정답 유형 분류보다 정확도가 높다는 결론을 얻었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 방법론 소개와 본 논문에서 제안하는 시스템 구조 및 정답 유형 분류 방법에 대해 설명한다. 4장에서는 실험과정에 대해 서술하고, 결과를 분석하였다. 마지막으로, 5장에서는 결론 및 향후 연구에 대해 기술하였다.

2. 관련 연구

정답 유형이라는 개념은 1999년 TREC(Text REtrieval Conference)-8에 출전했던, 질의 응답 시스템 LASSO에 의해 처음으로 도입되었다[1]. 이후 많은 질의 응답 시스템에서 정답 유형 분류 단계를 질의 응답 시스템 개발에서 활용하였다[1-8].

2001년 TREC-10에 출전했던 질의 응답 시스템 SiteQ[7]는 2단계 계층 구조를 가지는 정답 유형 분류 체계를 구성하였다. 또한 어휘 의미 패턴(Lexical-Semantic Patterns, LSP)을 이용하여 정답 유형을 결정하였다. 사용자의 질의를 미리 정의한 LSP에 대응시켜서 정답 유형을 결정하였다. 예를 들면,(%who)(%be)(@person)->PERSON이라는 LSP가 있다. “Who was president Cleveland's wife?"라는 예문은 (%who)(%be)(@person)라는 LSP에 대응하기 때문에 예문의 정답 유형은 PERSON이 된다. 이 방법론을 적용하기 위해, 기존 QA Track에 사용한 질문들과 웹 데이터를 수집하여 수작업으로 361개의 LSP를 구성하는 노력이 선행되었다. 2007년도에 TREC에 출전한 Ephyra라는 질의 응답 시스템은 154개의 정답 유형을 사용했으며,[6]과 마찬가지로 계층 구조를 가지는 정답 유형 분류 체계를 구성했다. TREC에서 사용한 질문들을 반영하여 정답 유형 클래스를 구성했다. 정답 유형 분류에 대해 비중 있게 다루고 있으며, 규칙 기반의 방법과 통계적 학습 모델 방법 두 가지를 하이브리드 하여 사용하였다[3].

ETRI에서는 [3]의 Ephyra 시스템과 마찬가지로, 구조적 자질 벡터 기계(structured Support Vector Machine)와 규칙에 기반한 방법을 하이브리드(Hybrid)하였다[5]. 통계 방법을 적용하기 위한 학습데이터를 제작하기 위해, 약 82000개 질문에 대해 정답 유형을 수동 태깅하는 노동력이 필요했다. 이 때 사용한 질문데이터는 국내 지식검색 사이트에서 추출한 것이다. 이처럼, 정답 유형이 태깅된 데이터를 구하는 것은 쉽지 않기에 자체적으로 수동 태깅을 한 경우가 대다수이다. 또한 정답 유형이 태깅된 데이터는 영어 외에 기타 언어에서는 양적인 측면에서 더욱 부족한 실정이므로, 영어 데이터를 번역하여 훈련 데이터로 사용한 경우도 있었다[2].

앞서 언급했던 패턴 매칭 방법이나 교사학습을 이용한 정답 유형 결정 방법은 다량의 수작업을 필요로 한다. 본 논문에서는 반교사 학습을 통해 정답 유형 분류에 필요한 노동력을 최소화하는 방법을 제안한다.

3. 반교사 정답 유형 추출 방법

3.1 LDA를 이용한 정답 유형 분류

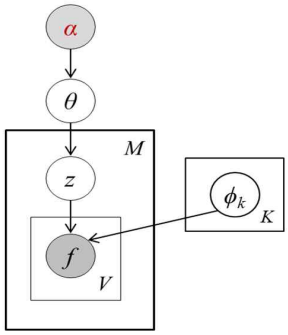

그림 1 정답 유형 분류를 위한 LDA의 도식

LDA는 2003년 M.Blei에 의해 제안된 비교사 학습(Unsupervised Learning)으로 텍스트 말뭉치와 같은 이산형 데이터(Discrete data) 집합에 대한 생성 확률 모델(Generative Probabilistic Model)이다. LDA는 베이지안 모델(Bayesian model)로써, 생성 확률 모델은 확률과 파라미터로부터 데이터가 생성된다는 관점을 나타낸 것이다. 정답 유형을 결정하기 위해서 LDA를 적용할 때는, 기존의 정답 유형 분류와 다르게, 정답 유형을 기준으로 클러스터가 생성된다고 본다. 즉, 정답 유형으로부터 여러 자질들로 표현되는 질의가 생성되는 것으로 볼 수 있다. 관련 연구로는 사용자의 발화 의도로부터 발화를 구성하는 단어들이 생성된다는 관점에서의 연구가 있었다[9].

정답 유형의 분포를 나타내는 θ는 하이퍼 파라미터(Hyper Parameter) α의 디리슐레 프로세스(Dirichlet Process)로부터 생성 된다. 상태(State),즉 클러스터에 해당하는 z는 정답 유형을 나타낸다. z는 θ의 다항 분포(Multinomial distribution)를 따른다. 각 문장에서 추출한 자질들을 나타낸 f는 정답 유형 z와 파라미터 ϕ_k의 다항분포에 의해 생성 된다. 본 논문에서는 일부 태깅된 데이터를 이용하여 Semi-Supervised LDA를 정답 유형 분류에 적용한다.

3.2 제안하는 질의 응답 시스템

본 논문에서는 DBpedia[1] 나 YAGO[2](Yet Another Great Ontology) 와 같이 구조화된 DB를 활용한 질의 응답 시스템을 제안한다. 위키피디아에서 구조적인 정보를 추출하여 공개적으로 제공하는 DBpedia를 활용하였고, 위키피디아와 워드넷(WordNet)에서 추출한 온톨로지인 YAGO도 DB로 활용할 계획이다. 뿐만 아니라 FreeBase[3]

1) http://dbpedia.org
2) http://www.mpi-inf.mpg.de/yago-naga/yago/
3) http://www.freebase.com/

나 기타 구조화된 DB를 추가하여 DB를 확장하는 연구를 진행하고 있다. 궁극적으로는 웹 기반 오픈 도메인 질의 응답 시스템으로 개발을 확장할 것이다. 본 논문에서 제안하는 질의 응답 시스템은 그림 2와 같다.

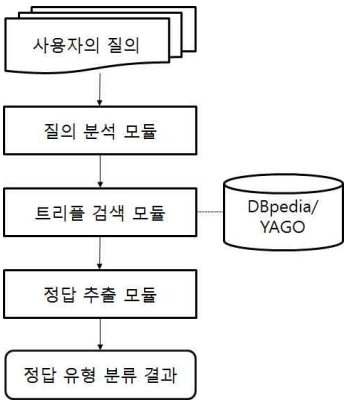

그림 2 제안하는 질의 응답 시스템 구조도

3.2.1 질의 분석 모듈

사용자의 질의에서 <Subject-Property-Object> 형태의 트리플을 추출한다. 사용자의 질의에서 사용자가 찾고자 하는 정보인 질문 초점(Question Focus)을 찾는다. 질문 초점은 예를 들어 "What is the **largest city** in Germany?" 라는 문장이 있을 때, Largest city가 된다. 질문 초점을 Subject또한 Object로 추출하고, 질문 초점의 유형으로, 정답 유형을 추출한다. 정답 유형을 결정하여 최종적으로 정답을 추출하는데 활용한다. Parser와 relation Extractor를 혼용하여 Property 와 Subject 또는 Object를 추출한다.

3.2.2 트리플 검색 모듈

사용자의 질의에서 <Subject-Property-Object> 형태의 트리플과 정답 유형을 추출하여 DBpedia 및 YAGO 와 같은 구조화된 DB에서 검색한다. Property와 Object 또는 Subject가 일치하고 정답 유형이 일치하거나, 정답 유형의 하위 유형(정답 유형이 사람일 때, 정답 유형 분류 체계가 사람-가수일 때, 검색 대상이 가수인 경우)이 일치하거나, 상위 유형(정답 유형이 가수일 때, 검색 대상이 사람인 경우)이 일치하는 트리플들을 검색한다.

3.2.3 정답 추출 모듈

정답 유형과 Subject, Property, Object 등을 이용하여 Scoring measure를 통해 순위를 정한다. 최종적으로 가장 정답일 확률이 높은 것을 정답으로 추출한다.

3.3 제안하는 정답 유형 분류

본 논문에서는 일부 태깅된 데이터를 이용한 Semi-Supervised LDA를 적용하여 정답 유형을 분류한다. 이를 통해 정답 유형 태깅에 대한 노동력 절감 효과와 수동 태깅에서 비롯된 모호성 문제를 해결할 수 있다. 본 논문에서 제안하는 정답 유형 분류는 그림 3과 같다.

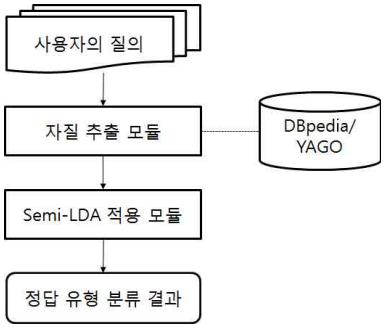

그림 3 제안하는 정답 유형 분류

3.3.1 자질 추출 모듈

LASSO 시스템에서 정답 유형은 의문사를 이용한 질문 유형(Question Type)과 질문 초점(Question Focus)을 통해 결정한다[1]. 따라서, 의문사와 질문 초점을 자질로 사용했다. 그 외에 자질로는 사용자 질의의 본동사(main verb)가 있다. 본동사는 사용자의 질의 의도를 반영하는 경우가 많기 때문에 자질로서 추출했으며, TREC에서 우수한 성능을 보인 Ephyra 시스템[3]에서도 추출하였다. 본동사 추출을 위해서 Stanford Parser[1]를 이용하였다. 또한 질의에서 의문사의 앞 뒤 어휘 정보를 이용하였다. 이를 통해 how many, how much 등을 구별할 수 있다. 그 외에 어휘 정보로는 다른 품사보다 문장에서 중요한 역할을 하는 명사, 동사에 해당하는 어휘 정보를 사용하였다. 또한 DBpedia를 활용하여 질의에 고유 명사가 있는 경우 DBpedia에서 제공하는 <개체명(Named-Entity),개체 유형(Named-Entity Type)> 데이터에서 검색하여 고유명사의 개체 유형 정보를 자질로 이용하였다.

3.3.2 Semi-LDA 적용 모듈

앞서, 추출된 자질을 바탕으로 Semi-Supervised LDA를 적용하여 정답 유형을 분류하였다. 정답 유형 분류는 구조화된 DB에서의 검색을 위해 위키피디아와 Wordnet을 연동한 YAGO의 분류 체계를 활용하였다. YAGO의 분류 체계는 약 1,700,000가지 개체들로 구성되어 있다. UIUC의 정답 유형은 이에 모두 1:1 대응 가능하다. 뿐만 아니라 UIUC의 other과 같은 유형은 더 세분화 가능하여 대응 가능하다. 일부 태깅된 데이터를 포함하여 클러스터링 하기 때문에 각각의 클러스터 ID에 해당하는 정답 유형이 정해져 있다. 정답 유형 분류에 대한 성능은 실험에서 기술하였다.

4. 실험

4.1 실험 설계

정답 유형 분류에서 많이 사용하고 있는 교사학습 방법과 본 논문에서 제안하는 반교사 학습 방법을 비교한다. 반교사 학습은 교사 학습과 다르게 일부 태깅 데이터를 통해 학습이 가능하다는 장점이 있음으로 이 부분을 검증한다.

교사 학습 방법으로 CRF(Conditional Random Fields)

1) http://nlp.stanford.edu/software/lex-parser.shtml

를 이용한다. 정답 유형 분류에 대한 CRF와 Semi-Supervised LDA의 정확도를 측정한다. 전체 학습 데이터에서 정답 유형이 태깅된 데이터의 비율에 따른 각 알고리즘의 성능을 측정한다. 랜덤으로 태깅 데이터를 추출하며, 교사 학습인 CRF와의 성능 비교를 위해 각 정답 유형이 1회 이상 훈련 데이터에 반영되도록 한다. 각각의 정답 유형에 대한 성능을 측정하여 성능이 높은 정답 유형과 성능이 낮은 정답 유형을 비교 분석한다.

4.2 데이터

UIUC(University of Illinois at Urban-Champaign)가 정답 유형 분류 실험에서 사용한 약 5500개 질의를 훈련 데이터로 사용하였다. 테스트 데이터로는 TREC 10에서 사용한 500개를 사용하였다. UIUC에서 정답 유형 분류 실험에 사용한 데이터는 기존의 연구들에서도 많이 사용한 공신력 있는 데이터이다[2]. TREC에서 제공하는 테스트 데이터도 질의 응답 시스템 관련 연구에서 성능 측정의 목적으로 활발히 활용하고 있다[2].

본 논문에서는 실험 결과를 실제 질의 응답 시스템 개발에 활용하는 것을 목적에 두고 있다. 따라서, 위키피디아 기반 질의 응답 시스템 제작을 위해 YAGO의 분류 체계와 UIUC의 정답 유형을 대응하는 작업을 선행하였다. 숫자나 날짜 등 어떤 글자도 YAGO 온톨로지에서 개체로 표현되어 있다. 따라서, YAGO의 분류 체계에 UIUC의 각 정답 유형이 모두 대응 가능함으로 UIUC의 50가지 정답 유형 분류 체계와 동일한 분류 체계를 이용하였다.

표 2 UIUC에서 제공하는 정답 유형 분류 체계

Coarse Class	Fine Classes
ABBREV.	abbreviation, expression
ENTITY	animal,body,color,creative,currency,disease,event,food,instrument,lang,letter,other,plant,product,religion,sport,substance,symbol,technique,term,vehicle,word
DESCRIPTION	definition,description,manner,reason
HUMAN	group,individual,title,description
LOCATION	city,country,mountain,other,state
NUMERIC	code,count,date,distance,money,order,other,period,percentage,speed,temp,volumesize,weight

4.3 실험 결과 및 분석

표 3은 훈련 데이터 중 정답 유형이 태깅된 데이터의 비율에 따라 Semi-Supervised LDA와 CRF의 정확도를 비교한 것이다. Semi-Supervised LDA는 정답 유형이 태깅된 데이터와 태깅되지 않은 데이터를 모두 훈련에 사용하였다. CRF는 훈련 데이터에 모두 정답 유형이 태깅되어야 하기 때문에, 태깅된 데이터만 훈련 데이터로 사용하였다. 표 3에 따르면, 태깅된 데이터가 10% 일 때, 즉 5000개 이상의 훈련 데이터 중에서 500개 이상의 태깅된 데이터를 포함하여 클러스터링 하였을 때, CRF보다 높은 성능을 보인다. 이 결과는 반교사 학습을 통해 정답 유형을 분류하는 것이 정답 유형 태깅 노동력을 줄일 수 있다는 가능성을 나타낸다. 25% 태깅된 경우는 거의 동일한 성능을 보이며, 50%, 60%, 75%, 100% 태깅 되었을 경우에는 CRF와 같거나 높은 성능을 보인다. 실험 결과에 대한 분석을 위해, 높은 정확도를 보인 정답 유형들과 낮은 정확도를 보인 정답 유형들을 분석하였다.

표 3 태깅된 데이터 비율에 따른 정확도 비교

Percentage	Semi-Supervised LDA	CRF
1%	0.036	0.098
10%	0.376	0.252
25%	0.422	0.444
50%	0.410	0.410
60%	0.528	0.450
75%	0.578	0.450
100%	0.630	0.454

Semi-Supervised LDA을 통해 정답 유형을 분류하였을 때, 수량(NUM_COUNT), 날짜(NUM_date), 사람(HUM_ind)에 해당하는 정확도가 높았다. 반면, 스포츠(ENTY_sport)나 화폐(ENTY_currency), 교통수단(ENTY_veh), 사건(ENTY_event) 등은 정확도가 낮았다.

정확도가 높았던 정답 유형들을 특징들을 분석한다. 첫째, 정확도가 가장 높았던 정답 유형인 수량은 "How many ~" 로 시작하는 경우가 대부분이라는 특징을 보였다. “How many” 또는 “How much" 는 정답 유형이 수량인 경우를 제외하고 거의 쓰이지 않은 어휘 정보이다. 둘째, 높은 정확도를 보였던 날짜 역시 주로 "When was" 로 시작되는 문장 유형이 많았고, 질문 초점인 “year"가 빈번하게 나타났다. 또한, 정답 유형이 날짜로 클러스터된 질의의 자질 중에 본동사가 was인 경우가 대부분이다. 셋째, 정답 유형이 사람인 경우도 의문사가 대부분 “who" 라는 특징이 두드러지며, 본동사로 "was" 나 "is"가 포함된 문장이 많다. 또한, 문장 내에서 고유 명사를 포함하는 경우가 많다. 뿐만 아니라, 전체 훈련 데이터에서 정답 유형이 사람인 질의가 차지하는 비중이 약 1/5 이상으로 높다.

정확도가 낮았던 정답 유형들 중 교통수단에 대한 정확도를 분석한다. 실제 정답 유형은 교통수단이지만, 정답 유형이 사람, 그룹, 날짜, 동물 등 다양한 정답 유형으로 클러스터링 된 경우가 많다. 예를 들어 정답 유형이 사람인 것에 클러스터링 된 경우는 “What was the name of the plane Lindbergh flew solo across the Atlantic ?”가 있는데 본동사와 고유 명사와 같은 자질에 영향을 받았을 것으로 분석된다. 클러스터링은 “빈익빈 부익부” 의 성격을 갖기 때문에, 새로운 데이터는 기존에 형성된 클러스터의 크기에 비례하여 클러스터링 된다. 즉, 새로운 데이터는 크기가 큰 클러스터에 속할 확률이 크다[10]. 실제로 크기가 큰 클러스터에 다른 정답 유형을 가지는 질의가 속한 경우는 많았다. 하지만, 다른 정답 유형을 가지는 질의가 교통수단 클러스터처럼 작은 클러스터에 속한 경우는 거의 없었다. 즉, 훈련 데

이터에서 큰 비중을 차지하는 사람과 같은 정답 유형은 사람이라는 클러스터에 다른 정답 유형들이 함께 포함되어 성능이 떨어지는 경향이 있다. 반면에, 훈련 데이터에서 작은 비중을 차지하는 교통수단과 같은 정답 유형은 교통수단에 속하는 질의가 다른 클러스터들에 속하게 되어 정확도가 떨어지는 경향을 보였다. 아래 그림 4는 훈련 데이터에서 큰 비중을 차지하는 정답 유형이 정확도가 높은 예들을 보여준다. 가로축은 각 정답 유형이 훈련 데이터에서 나타나는 빈도수이며 세로축은 각 정답 유형의 정확도이다. 훈련 데이터의 태깅 비율이 달라도 데이터의 분포는 그림 4와 비슷한 양상을 띤다.

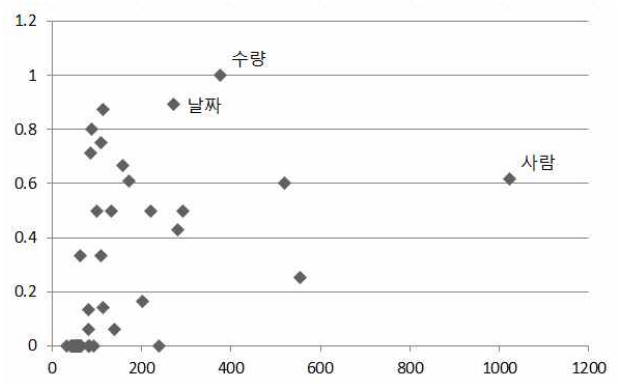

그림 4 50% 태깅 데이터로 훈련했을 때, 훈련 데이터에서 차지하는 비중에 따른 정확도 분포

5. 결론 및 향후 연구

본 논문의 실험을 통해 동일한 양의 태깅 데이터가 주어졌을 때, Semi-supervised LDA가 CRF보다 대체적으로 높은 성능을 보이는 것을 확인할 수 있었다. Semi-supervised LDA는 태깅된 데이터와 태깅이 안된 데이터를 함께 이용할 수 있지만, CRF는 태깅된 데이터만 학습에 이용할 수 있다. 따라서, 반교사 기반의 정답 유형 분류 방법을 통해 정답 유형 태깅 노동력을 줄일 수 있다는 결론을 얻었다.

하지만 임의의 질의 말뭉치들을 정답 유형 별로 클러스터링 할 때, 전체 데이터에서 차지하는 비중이 작은 클러스터들의 정확도를 향상시키는 연구가 필요하다.

또한 웹에서 수집한 질문의 경우 정답 유형이 정해져 있지 않다. 이러한 대용량의 데이터에 정답 유형을 분류하기 위해서 계층적 드리쉴레 프로세스(Hierarchical Dirichlet Process ,HDP)를 이용한 연구를 진행 중이다. HDP를 통해 분류된 각각의 클러스터들을 YAGO의 분류 체계에 대응시키는 연구를 진행할 것이다.

뿐만 아니라, 본 논문의 연구결과를 활용하여, 정답 유형 레이블이 부족한 한국어 질의 데이터를 자동 레이블링할 것이다. 이를 바탕으로, 한국어 질의 응답 시스템 개발에 대한 연구를 진행할 것이다.

*본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10044508 , 비기호적 기법 기반 인간 모사형 자가학습 지능 원천기술 개발]

참고문헌

[1] Dan Moldovan et al., "LASSO: A Tool for Surfing the Answer Net," TREC, Vol.8, p.65-73, 1999

[2] Anne-Laure Ligozat, "Question Classification Transfer," Proceedings of the Association for Computational Linguistics, p.429-433, 2013

[3] Nico Schlaefer et al., "Semantic Extensions of the Ephyra QA System for TREC 2007," TREC, 2007

[4] Xin Li et al., "Learning question classifiers," Proceedings of the international conference on Computational linguistics", Vol.1, p.1-7, 2002

[5] 허정 외, "오픈 도메인 질의응답을 위한 검색문서 제약 및 정답유형 분류기술," 정보과학회논문지:소프트웨어 및 응용, 제39권, 제2호, 2012

[6] Xin Li et al., "The role of semantic information,"Natural Language Engineering, Vol.12, no.3, p.229-249, 2006

[7] Gary Geunbae Lee et al., SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP," TREC, 2001

[8] 송일현 외 , 실시간 검색어를 이용한 주제어 기반의 질의 응답 시스템, 제 23회 한글 및 한국어 정보처리 학술대회, 2011

[9] Donghyeon Lee et al., "Unsupervised modeling of user actions in a dialog corpus," Proceeding of the IEEE international conference on acoustics, speech, and signal processing, 2012

[10] David Blei et al., "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," Journal of the ACM, Vol.57, no.2 , 2010