URI 중의성 해소 및 오류 감소를 위한 LDA 기반 접근법

김지성○, 김영식, 함영균, 황도삼, 최기선
한국과학기술원, 한국과학기술원, 한국과학기술원, 영남대학교, 한국과학기술원
jiseong@kaist.ac.kr, twilight@kaist.ac.kr, hahmyg@kaist.ac.kr, dshwang@yu.ac.kr, kschoi@kaist.edu

LDA-based Approach for URI Disambiguation and Error Reduction

Jiseong Kim○, Youngsik Kim, Younggyun Hahm, Dosam Hwang, Key-Sun Choi
KAIST, KAIST, KAIST, Yeungnam University, KAIST

요 약

URI 중의성 해소 문제는 주어진 문서 내의 특정 단어에 연결 가능한 여러 URI가 주어졌을 때 진짜 URI 하나를 선택해내는 문제라고 할 수 있다. 이 문제는 다양한 해결법들이 존재할 수 있지만 기존에 연구된 문서의 문맥 간 유사도를 이용하여 해결하는 방법을 본 논문에서는 사용한다. 문맥 간 유사도를 이용하는 방법은 영어 디비피디아 URI spotting에서 TF*ICF방법으로 이미 연구가 되어있다. 본 논문에서는 Latent Dirichlet Allocation을 이용하여 URI 중의성 해소 문제를 다룰 것이며 그 범위를 한국어 디비피디아로 한정한다. 새로 제안하는 방법이 URI 중의성 해소 문제를 얼마나 잘 해결하며, 기존의 연구와 비교하여 얼마나 향상될 수 있는지를 분석한다. 또한 기존의 방법과 새로 제안한 방법 각자가 고유하게 풀 수 있는 문제가 존재함을 보이고, 두 방법을 병합하였을 때 보다 높은 성능에 도달할 수 있음을 전망한다.

주제어: URI 중의성 해소, 토픽 모델, 한국어 디비피디아, 링크드 데이터

1. 서론

임의의 문장이 주어졌을 때 여기서 개체가 될 만한 단어 혹은 표층형(surface form)을 식별하고(개체 경계 식별), 이 표층형에 연결될 수 있는 여러 개체 중 하나를 골라내는 문제(개체 중의성 해소)는 이미 많이 알려져 있고 연구되어 왔다. 대표적인 예로 영어 디비피디아 URI spotting[1] 연구가 있다.

개체명 인식 문제에서 개체는 다양하게 정의 가능하다. 사람, 동물, 장소 명 등등이 될 수 있으며 인터넷 상에서의 웹 페이지 혹은 기타 자원의 URI가 될 수도 있다. 이는 개체명 인식 문제를 풀고자 하는 사람이 정의하기 나름이다. URI 중의성 해소 문제는 개체명 인식 문제의 일부분의 특수한 경우라고 볼 수 있다.

개체명 인식 문제의 어려운 점 중 하나는 단어의 중의성 때문에 발생한다. 식별된 표층형은 포함되어 있는 글의 문맥에 따라 다양한 뜻을 가질 수 있으며, 이는 하나의 표층형에 연결 가능한 개체가 1개가 아닌 여러 개가 될 수 있음을 나타낸다. 문맥을 고려했을 때, 주어진 표층형에 가장 적합한 개체를 골라내는 것이 개체 중의성 해소 문제이며, 특별히 개체가 웹페이지와 같은 URI를 갖는 자원(resource)일 경우 이를 URI 중의성 해소 문제라고 부른다. 대표적인 예로 영어 디비피디아 URI spotting 같은 경우 임의의 문장에서 개체가 될 만한 표층형을 식별하는 개체 경계 식별 문제와 표층형과 연관있는 디비피디아 자원의 URI를 선택하는 URI 중의성 해소 문제를 다루고 있다. 이 논문에서는 한국어 디비피디아를 대상으로한 URI spotting 문제 중 개체 경계 식별 문제는 다루지 않으며 오직 URI 중의성 해소 문제에 초점을 둔다.

영어 디비피디아에서는 URI 중의성 해소 문제를 풀기 위해 TF*ICF(Term Frequency * Inverse Candidate Frequency)[1]를 사용하였다. 본 논문에서는 한국어 디비피디아를 대상으로 LDA(Latent Dirichlet Allocation)[2]를 사용하여 URI 중의성 해소 문제를 해결하는 방법을 제안한다. 또한 영어 디비피디아 URI spotting에서 사용되었던 베이스라인 및 TF*ICF를 이용한 방법과 본 논문에서 제안하는 LDA 토픽모델을 사용한 방법을 한국어 디비피디아 URI 중의성 해소 문제에 적용하여 비교하였으며, 각자 방법이 잘 풀 수 있는 경우가 있음을 보이고, 이에 따라 각 방법이 상호 보완적 관계가 될 수 있다는 가능성을 전망한다.

2. 관련 연구

2.1 문서의 유사도 계산

문서의 내용 간 유사도를 계산 하는 대표적인 방법에는 cosine 유사도(similarity)가 있다. 문서 집합에서 중요한 키워드를 추출하여 순서에 상관없이 하나의 벡터(bag-of-words)로 표현하고, 각 문서마다 키워드들의 빈도수(Term Frequency)를 측정하여 키워드에 해당하는 벡터의 원소에 저장한다. 각 벡터의 원소마다 IDF(Inverse Document Frequency)를 곱해준 값이 최종적으로 그 문서를 대표하게 된다. 두 문서의 내용 간 유사도는 두 문서를 대표하는 벡터 간의 cosine 유사도를 이용하여 계산할 수 있다. 디비피디아 URI 중의성 해소[1]에서는 유사도를 계산하기 위해 TF*IDF 대신 TF*ICF(Term Frequency * Inverse Candidate Frequency)를 사용하였다. TF*ICF는 모든 문서가 아닌 특정 표층형의 후보문서들(Candidates)을 대상으로 하는 ICF 값을 사용한다. ICF 값은 다음과 같은 식에 의해 구할 수 있다.

$$ICF(w) = \log \frac{|\text{표층형 } s \text{의 후보문서집합 } S|}{|\text{단어 } w \text{가 등장하는 } S \text{의 부분집합}|}$$

수식 1 단어에 대한 ICF 값을 구하는 식 : 어떤 문서 d에 속한 표층형 s의 URI 후보문서집합 S가 주어졌을 때 d에 속한 어떤 단어 w의 ICF값을 구하는 역할을 한다.

디비피디아 URI 중의성 해소 문제에서는 후보문서들 중 하나를 선별하는 것이 중요하기 때문에, 후보문서들 간 빈번히 나오는 단어는 식별하는 기준으로 쓰기에는 부적합하다. 따라서 IDF가 아닌 ICF를 사용하여 이런 단어들의 중요도를 낮추어 준다.

2.2 Latent Dirichlet Allocation (LDA)

LDA는 단어 집합과 같은 이산형 데이터를 모델링하기 위한 generative probabilistic model이다. 정보 검색(Information retrieval) 분야에서는 여러 문서들의 토픽 분석을 위해 많이 사용된다. LDA에서 한 문서는 여러 토픽들의 혼합으로 구성되어 있으며 이에 대한 확률분포를 갖는다고 가정한다. 또한 하나의 토픽은 여러 단어들의 혼합으로 구성되며 이에 대한 확률분포를 갖는다고 가정한다. LDA는 훈련에 필요한 문서들과 토픽 개수가 주어지면 문서 내 혹은 문서 간의 단어 빈도수를 분석하여 문서에 대한 토픽분포, 토픽에 대한 단어분포를 추정해낸다. 현실적으로 정확한 값을 추정해내기는 어려우며, 이를 위해 Laplace 근사법, variational 근사법, Markov chain Monte carlo[2]등과 같은 근사하게 추정하는 기법들이 연구되어 있다. 학습 후 얻은 토픽 모델을 이용하여 학습 데이터가 아닌 새로운 데이터가 들어왔을 때 이 데이터의 토픽분포 또한 추정[3]해 낼 수 있다.

2.3 LDA 기반 문서의 유사도 계산

LDA를 통해 구한 문서의 토픽분포는 그 문서를 대표하게 된다. 문서 간 유사도를 구하기 위해서는 그 문서들을 대표하는 토픽분포 사이의 유사도[3]를 구하면 된다. 확률분포 간 유사도를 측정하기 위해서는 KL divergence(Kullback-Leibler divergence)[4]를 사용할 수 있다. 다음과 같은 식에 의해 두 이산확률분포 P와 Q 사이의 유사도를 구할 수 있다.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$

단, 위의 함수는 인자 순서에 대칭적이지 않으므로 본 논문에서는 다음과 같은 대칭적인 KL divergence를 사용한다.

$$D_{\mathrm{KL}}(P\|Q) - D_{\mathrm{KL}}(Q\|P)$$

3. 접근법

3.1 URI의 정의

본 논문에서 언급하는 URI는 모든 한글 위키피디아 문서들의 URI로 한정한다. 단 다음과 같은 두 가지 예외 경우에 대해서는 다르게 정의한다.

- 리다이렉션 : 일부 위키피디아 URI는 고유 문서를 가지고 있지 않고, 자동으로 다른 URI로 리다이렉션 된다. 따라서 이런 URI는 리다이렉션 후의 문서의 URI로 정의한다.
- 동음이의: 일부 위키피디아 URI의 문서는 특정 개체를 의미하지 않고 같은 낱말을 공유하는 동음이의어들의 링크만 가진 경우가 있다. 이 URI들은 특정 문서를 가리키고 있지 않으므로 URI로 인정하지 않았다.

3.2 문제 정의

개체명 인식 문제는 다음과 같은 과정에 의해 해결될 수 있다.

1. 개체 경계 식별 : 개체가 될 수 있는 표층형을 식별한다.
2. 개체 후보 선별 : 식별된 표층형과 연결될 수 있는 가능한 모든 개체 후보들을 고른다.
3. 개체 중의성 해소 : 개체 후보들 중 식별된 표층형에 적합한 하나의 개체를 선택하여 표층형에 부여한다.

본 논문에서는 URI 중의성 해소 문제를 풀기 위한 방법에 주안점을 두고 있으므로 1, 2번 과정은 다루지 않고 오직 3번 과정에 집중한다. 따라서 처리해야할 입력은 임의의 문장뿐만이 아니라 문장에서 추출한 표층형들과 각각의 표층형에 대한 URI 후보들까지 포함하며, 각각의 표층형에 대해 올바른 URI 후보를 하나 선택하여 해당 표층형에 부여하는 것이 본 논문에서 중심적으로 다룰 문제 처리 과정이자 출력이 된다.

3.3 문서의 유사도를 이용한 URI 중의성 해소

개체 중의성 해소 문제는 하나의 표층형과 이에 해당하는 여러 개체 후보 중 가장 적합한 것을 고른다는 점에서 WSD(Word Sense Disambiguation) 문제와 유사하다. WSD 문제 관점에 비추어 표층형은 동음이의어라고 볼 수 있으며 개체 후보들은 동음이의어에 딸린 여러 뜻에 해당한다고 볼 수 있다. 동음이의어의 뜻은 그 단어를 포함하는 주변 문장들의 문맥에 의해 높은 확률로 결정된다는 연구 결과[5]가 존재한다. URI 중의성 해소 문제를 WSD 문제를 이용하여 재조명하면 다음과 같이 생각할 수 있다.

1. 동음이의어는 URI 중의성 해소 문제에서의 표층형이라 생각할 수 있다.
2. 동음이의어에 속하는 여러 뜻들은 URI 중의성 해소 문제에서 URI 후보들이라 생각할 수 있다.

3. 어떤 문장 A에 포함된 동음이의어 w의 뜻은 A와 비슷한 문맥을 갖는 어떤 문장 B에 포함된 동음이의어 w가 갖는 뜻과 높은 확률로 일치한다.[5]
4. 1, 2, 3번 항목에 비추어, 어떤 문서 D에 포함된 표층형 s의 URI는 비슷한 문맥을 갖는 어떤 문서 E에 포함된 표층형 s의 URI와 같을 확률이 높다.

위 목록에서 1, 2, 4번이 성립될 수 있는 이유는 디비피디아의 URI에 해당하는 각각의 자원들은 일반적으로 어떤 단어에 대한 의미 혹은 설명을 포함하는 위키피디아의 문서들과 상응하기 때문이다. 즉 위키피디아를 일종의 사전으로 생각한다면, 각 표층형에 이 사전에 포함된 여러 가능한 해석(문서) 중 하나를 부여하는 것이라 볼 수 있다. 본 논문에서는 위와 같은 가설을 염두에 두고, 표층형을 포함하는 문서와 그 표층형과 연결된 URI의 링크를 갖는 위키피디아 문서들 간의 문맥 간 유사도를 이용하여 URI 중의성 해소 문제를 해결하고자 한다.

3.4 LDA 기반 URI 중의성 해소

본 논문에서는 URI 중의성 해소 문제를 해결하기 위해 문서의 문맥 간 유사도를 이용하며, 이를 위해 LDA를 사용하는 방법을 제안한다. LDA는 단어 간 동시등장(co-occurence)을 반영하여 토픽모델을 생성해낸다. 한 문서에서 자주 같이 등장하는 단어들은 같은 토픽에 묶일 확률이 높다. 이 때문에 디비피디아 URI spotting에서의 TF*ICF는 측정할 수 없는 문맥 간 유사도를 LDA는 측정할 수 있는 경우가 발생한다. 예를 들면 '빅데이터'와 '데이터마이닝'은 비슷한 분야에서 사용되는 연관성이 높은 단어들이다. 만약 어떤 두 문서 A, B가 같은 어휘를 전혀 사용하지 않으며 문서 A에는 '빅데이터'가, 문서 B에는 '데이터마이닝'이 다수 포함된다고 생각해보자. TF*ICF를 이용한 측정에서는 두 문서는 전혀 유사하지 않다고 결과가 나올 것이다. 반면 LDA를 이용하면 비슷한 토픽에 할당된 연관성이 높은 두 단어에 의해 어느 정도 유사하다는 결과가 나올 것이다. 본 논문에서는 이와 같은 경우들이 한국어 디비피디아의 URI 중의성 해소 문제를 대상으로 얼마나 자주 등장하며, 이를 이용하여 LDA가 TF*ICF에 비해 얼마나 성능 향상을 보일 수 있는지를 보이려 한다. 또한 반대로 TF*ICF가 LDA보다 더 나은 결과를 보이는 경우가 있음을 보이며, TF*ICF와 LDA가 상호보완적 관계가 될 수 있다는 가능성을 전망한다.

4. 실험

4.1 데이터 집합

실험에 사용한 문서 집합은 2014년 1월 26일자 한국어 위키피디아 문서 263,501개를 사용하였다. 각각의 위키피디아 문서에서 위키 문법에 속하는 구문 및 키워드를 모두 제거하여 제목과 내용으로만 구성된 문장으로 변환하였다. 변환 된 문장에서 표층형만을 식별하여 모아둔 것이 실험에서 사용할 데이터 집합의 후처리된 하나의 문서이다.

정답 데이터는 3명의 참여자에 의해서 수동으로 만들어졌다. 55개의 문서 및 이로부터 추출된 표층형 집합과 이에 대한 URI 후보들을 보고 3명의 참여자가 수동으로 정답을 골라내어 정답 데이터를 작성하였다.

LDA를 훈련하기 위한 데이터는 다음과 같이 두 가지 방식으로 구성된다.

- 식별된 표층형으로만 구성된 모든 위키피디아 문서들
- 한국어 형태소 분석기를 이용해 접두사, 접미사, 복합명사등을 고려하여 추출해낸 명사들로 구성된 모든 위키피디아 문서들

위 두 가지 방식에 대한 실험 결과(Precision, Recall, F1 score)를 각각 **LDA-surfaces**와 **LDA-nouns**로 명명한다.

4.2 URI 중의성 해소 알고리즘

실험에 사용한 중의성 해소 알고리즘은 랜덤 베이스라인, 빈도기반 베이스라인, TF*ICF, LDA이다. LDA를 제외한 3개의 방법은 디비피디아 URI spotting에서 사용한 방법을 그대로 가져온 것이다. 이를 이용해 한국어 디비피디아 URI spotting에 적용했을 때 어떤 결과를 보일지를 비교해 볼 수 있다. 또한 이 결과와 LDA의 결과를 비교함으로써 새로 제안하는 방법이 풀 수 있는 문제, 한계점 등을 조명할 수 있다. 다음은 각각의 알고리즘에 대한 실험 방법을 설명한다.

랜덤 베이스라인 이 베이스라인은 하나의 표층형과 선택해야 할 URI 후보들이 주어졌을 때, 무작위로 선택을 하는 방식이다. 즉 하나의 문서에서 등장하는 같은 표층형에 대해 다른 URI가 부여될 수 있다. 이 베이스라인의 목적은 URI의 중의성의 정도를 측정하는데 있다. 중의성이 높을수록 선택해야 할 URI 후보들은 많아지고 이로 인해 자연히 랜덤 베이스라인의 성능은 떨어질 것이다.

빈도기반 베이스라인 이 베이스라인은 하나의 표층형에 대해 선택해야 할 URI 후보들 중에서, 위키피디아 전체를 통틀어 그 표층형과 연결되었던 적이 가장 많은 URI 후보를 선택하는 방식이다. 확률적으로 높은 가능성을 갖고 정답 URI를 선택하지만, 문맥 정보를 고려하지 않기 때문에 틀린 URI를 고를 수도 있다.

TF*ICF 이 방식은 문서의 단어 빈도수를 하나의 벡터에 저장하고 각 원소마다 TF*ICF값을 가중치로 곱해준 최종 결과 벡터를 문서의 대표로 사용하고 이를 이용하여 문서 간 유사도를 측정한다. 이 때 단어의 단위는 식별가능한 표층형 혹은 형태소 분석기를 이용하여 식별되는 명사 두 가지를 사용하였다. 따라서 TF*ICF를 이용한 실험 결과는 LDA와 마찬가지로 두 가지가 되며 각각을 **TF*ICF-surfaces**와 **TF*ICF-nouns**로 명명한다.

구체적인 진행 과정은 다음과 같다. 문서 D의 표층형 s에 URI를 부여한다고 가정하자. 모든 위키피디아 문서

중에서 URI 중의성 해소를 하고자 하는 링크가 걸린 표층형 s를 포함하는 모든 문서를 추출한다. 추출된 각각의 문서는 저마다 다른 URI와 연결된 s를 포함하고 있을 것이다. 추출한 모든 문서와 URI를 달고자 하는 표층형 s을 포함하는 문서 D 간의 문맥 간 유사도를 측정한다. 이때 cosine 유사도를 이용하여 측정한다. 가장 유사한 문서에 포함된 표층형 s에 부여된 URI를 우리가 달고자 하는 문서 D의 표층형 s에 부여한다.

LDA 이 방식은 TF*ICF 방식과 처리과정이 대부분 같다. 한 가지 다른 점이라면 하나의 문서를 벡터와 TF*ICF로 표현을 하는 것이 아니라, 그 문서의 토픽분포로 표현을 한다. 또한 토픽분포 간 유사도를 측정할 때 KL divergence를 이용하여 계산한다. 문서의 토픽분포를 알아내려면 선행 처리 과정으로 LDA의 학습이 필요하다. 본 논문에서는 LDA를 학습시키기 위해 mallet 2.0.7 프레임워크[6]를 이용하였다. LDA를 훈련시키기 위한 토픽 개수는 300개로 설정하였다. 토픽 개수는 100개부터 500개까지 100개 단위로 바꾸어 가며 가장 나은 성능을 보이는 토픽을 고른 것이다. LDA를 위한 두 개의 학습 데이터를 이용해서 LDA를 각각 따로 학습시키고 두 개의 실험 결과를 산출하였다.

4.3 성능 측정 방법

본 논문에서는 성능을 측정하기 위해 CoNLL-2003 shared task[7]에 나온 성능 측정 방식(5-fold cross-validation)을 이용하였다. 정답 데이터 집합에서 정답을 제거한 표층형만을 추출하여 다섯 개의 균등한 집합으로 쪼개고, 이에 대해 4.2에서 언급한 각 알고리즘을 사용하여 저마다의 정답을 구한다. 이를 정답 데이터와 비교하여 5개의 F1 score를 구하고 평균을 내어 최종적인 성능을 구하였다.

4.4 실험 결과 및 고찰

표 1 URI 중의성 해소 알고리즘의 성능

알고리즘	Precision	Recall	F1 score
랜덤 베이스라인	61.55 %	60.15 %	60.84 %
빈도 베이스라인	90.57 %	88.49 %	89.52 %
TF*ICF-nouns	91.59 %	89.50 %	90.53 %
TF*ICF-surfaces	91.92 %	89.82 %	90.86 %
LDA-nouns	93.20 %	91.24 %	92.21 %
LDA-surfaces	92.70 %	90.75 %	91.71 %

실험 결과는 표 1에서 볼 수 있다. 베이스라인에 비해 TF*ICF와 LDA를 사용한 방법이 조금 더 높게 나왔다. 한국어 디비피디아 URI 중의성 해소의 빈도기반 베이스라인의 경우 영어 디비피디아의 경우보다 상당히 높게 측정되었다. 이는 한국어 위키피디아의 특징인 듯 하며, 대부분의 표층형은 연결 가능한 URI 후보 중 자주 연결되는 하나의 URI로만 매우 치우쳐서 연결된다고 볼 수 있다. 예를 들면 '대한민국' 이라는 표층형에 대해 '대한민국' 이라는 문서는 27,544번 연결된 것에 비해 그 다음으로 자주 연결된 문서는 '대한민국 축구 국가대표팀' 으로 153번이 연결되었으며, 모든 다른 약 20개의 후보문서들은 이보다 훨씬 적은 10개 이하의 연결 횟수를 보였다. 즉 '대한민국' 표층형이 하나의 문서로만 매우 치우쳐 연결되었으며, 대부분의 표층형도 이와 같은 양상을 보여 빈도기반 베이스라인의 성능이 매우 높은 것으로 보인다. 다시 말하면, 한국어 디비피디아 URI 중의성 해소 문제에서 대부분의 경우는 연결빈도를 이용하여 쉽게 해결 가능하며, 소수의 경우에만 문맥정보를 고려하여 URI를 선택해야 한다. 이런 이유 때문에 빈도 베이스라인의 성능과 TF*ICF, LDA의 성능 간 차이가 그렇게 크지 않게 나온 것으로 보인다.

TF*ICF와 LDA사이에도 성능 차이가 존재한다. 비록 그렇게 큰 수치는 아니지만 문맥 정보를 특별히 필요로 하는 문제가 그다지 많지 않았다는 점을 고려했을 때, 90% 이상의 성능 구간에서 약 1.35%가량의 향상은 어느 정도 LDA가 TF*ICF보다 문맥정보를 잘 식별함을 보여준다. 표 2는 각 알고리즘 별로 가장 성능이 좋은 빈도기반 베이스라인, TF*ICF-surfaces, LDA-nouns 각각이 다른 알고리즘이 틀린 것을 맞춘 경우의 전체 문제에 대한 비율을 나타낸다. LDA기반 방법이 TF*ICF방법에 비해 약 1.8배 가량 많이 맞추는 것을 볼 수 있다. 표 2를 보면 알 수 있듯이 TF*ICF방법이 잘 맞추는 경우가 있으며, LDA가 잘 맞추는 경우가 존재한다. 표 1과 표 2의 내용을 종합해보면 대부분의 경우(약 90%정도) 풀기 쉬운 경우에 대하여 TF*ICF와 LDA를 사용한 두 방법의 URI 중의성 해소 결과는 같지만 그 외의 경우 각 방법만이 해결할 수 있는 문제의 크기가 LDA가 TF*ICF보다 약 2배정도 크다. 표 2를 통해 더 확인할 수 있는 점은 한 방법이 다른 방법을 완전히 커버하는 것이 아닌 각 방법 고유의 잘 풀 수 있는 영역이 존재한다는 것이다. 이는 두 방법의 장점을 잘 살리는 방식으로 병합할 수 있다면 서로가 놓치는 부분을 보완하여 표 1에 나온 최고 성능보다 더 높은 F1 score을 달성할 수 있음을 내포한다. 즉 TF*ICF 방법과 LDA 방법의 병합 가능성을 전망할 수 있다.

표 2 각 알고리즘이 고유하게 맞춘 URI 중의성 해소 문제와 전체 문제에 대한 비율 : 1행은 빈도베이스라인은 맞추었지만 다른 두 방법이 맞추지 못한 정도를 나타낸다. 2행은 TF*ICF방법은 맞추었지만 LDA방법은 맞추지 못한 정도를, 3행은 그 반대를 나타낸다.

알고리즘	비율
빈도 베이스라인	0.86 %
TF*ICF-surfaces	1.87 %
LDA-nouns	3.27 %

5. 결론 및 추후 연구

본 논문에서는 URI 중의성 해소 문제에 있어서 기존의 TF*ICF 방법에 비해 LDA를 사용하는 방법이 어느 정도의 성능을 보이며 오류 감소에 얼마나 기여할 수 있는지를 살펴보았다. 비록 그 범위를 한국어 위키피디아라는 특

수한 경우로 한정했지만, 풍부한 어휘와 다른 말로 바꾸어 표현하는 것이 활발한 일반 인터넷 상의 문서들을 대상으로 할 경우 그 결과의 차이는 더욱 명확해질 것으로 기대한다.

본 논문에서 LDA를 훈련시키는데 사용한 토픽 개수는 300개이다. 단 이 결과는 토픽개수를 100개단위로 바꾸어가며 경험적으로 그나마 나은 것을 골라낸 것이지 최적의 토픽 개수는 아니다. 이를 보완하기 위해 추후 연구에서는 최적의 토픽 개수까지 추정하는 HDP-LDA[8]를 사용하여 문서 문맥 간 유사도 측정 분별력을 높일 계획이다.

LDA의 훈련 데이터의 질도 향상시킬 필요가 있다. 본 논문에서 사용한 LDA 훈련 데이터는 불용어가 제거되지 않은 채로 사용되었다. 본 논문에는 싣지 않았지만 하나의 단어가 여러 토픽에 빈번히 출연하는 것이 관찰되었다. 이는 토픽모델의 질을 떨어뜨려 문서의 문맥 간 유사도 측정의 분별력을 하락시킨다. LDA 훈련 데이터를 정제하여 다시금 실험을 진행할 경우 조금 더 높은 성능이 나올 것으로 기대한다.

이 논문에서는 TF*ICF, LDA 각 방법이 잘 푸는 문제가 존재함을 수치로써 보였지만, 이 수치 자체는 어떻게 TF*ICF와 LDA가 병합되어 사용될 수 있을지에 대한 정보를 제공하지 않는다. 보다 나아가 TF*ICF와 LDA가 잘 맞추는 문제들에 대한 분석이 필요하며, 분석이 명확해질 경우 두 방법을 병합할 수 있는 길을 모색할 수 있을 것이다. 추후에 정답 데이터를 위키피디아가 아닌 보다 일반적인 문서들로 풍부하게 확보하고, 그 결과를 분석하여, TF*ICF와 LDA가 병합 가능한지에 대해 모색할 계획이다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

참고문헌

[1] Mendes, Pablo N., et al., "DBpedia spotlight: shedding light on the web of documents", Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011.

[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[3] H Misra, O Cappe, and F Yvon. 2008. Using LDA to detect semantically incoherent documents. In Proc. of CoNLL 2008, pages 41-48, Manchester, England.

[4] Kullback, S. and Leibler, R.A., On information and sufficiency. Ann. Math Statist. 12 (1951) 79-86.

[5] Gale, William A., Kenneth W. Church, and David Yarowsky. 1992b. One sense per discourse. In Proceedings of the DARPA Speech and Natural Language Workshop

[6] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[7] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-147. DOI=10.3115/1119176.1119195 http://dx.doi.org/10.3115/1119176.1119195

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. In Proc. NIPS, 2004.

[9] 배덕호, 엄태환, 윤석호, 박정, 김상욱, "LDA를 이용한 논문 유사도 계산 방안의 성능 평가", [1] 한국통신학회 종합 학술 발표회 논문집 (동계) 2013, 2013.1, 356-357.